Requested Patent: JP11282636A

Title:

HIGH DENSITY RAID SUBSYSTEM WITH HIGHLY INTEGRATED CONTROLLER ;

Abstracted Patent: US6188571 ;

Publication Date: 2001-02-13 ;

Inventor(s):

BASCO JOSE PLATON (US); ROGANTI ADRIANO (US); WILLE THOMAS (US); SMITH RONALD BRUCE (US) ;

Applicant(s): AIWA RAID TECHNOLOGY INC (US) ;

Application Number: US19970963841 19971103 ;

Priority Number(s): US19970963841 19971103 ;

IPC Classification: G06F1/16 ;

Equivalents: ;

ABSTRACT:

The present invention provides a method and apparatus for a mass storage subsystem such as a RAID array. The invention includes a housing which defines first and second cavities with the first cavity housing an array controller such as a RAID controller. The second cavity houses a plurality of substantially conventional IDE drives conforming to the 3.5" form factor. The array is configured to maximize cooling of the array controller and the drives within the extremely small space defined by the housing

(19)日本国特許庁（ＪＰ）　(12)公開特許公報（A）　(11)特許出願公開番号

特開平11－282636

(43)公開日　平成11年(1999)10月15日

(51)Int.Cl.[6]		識別記号	FI		
G06F	3/06	540	G06F	3/06	540
G11B	33/02	301	G11B	33/02	301F
	33/14	503		33/14	503

審査請求　未請求　請求項の数1　OL　外国語出願　（全 55 頁）

(21)出願番号　特願平10－350665

(22)出願日　平成10年(1998)11月４日

(31)優先権主張番号　９６３８４１

(32)優先日　1997年11月３日

(33)優先権主張国　米国（US）

(71)出願人　000000491
アイワ株式会社
東京都台東区池之端１丁目２番11号

(72)発明者　アドリアーノ　ロガンティ
アメリカ合衆国　フロリダ州　マーゲイト，エヌ．ダブリュ．　サーティファースト　ストリート　5320

(72)発明者　ロナルド　ブルース　スミス
アメリカ合衆国　フロリダ州　ウエリントン，マイスティック　ウェイ　1137

(74)代理人　弁理士　浅村　皓　（外３名）

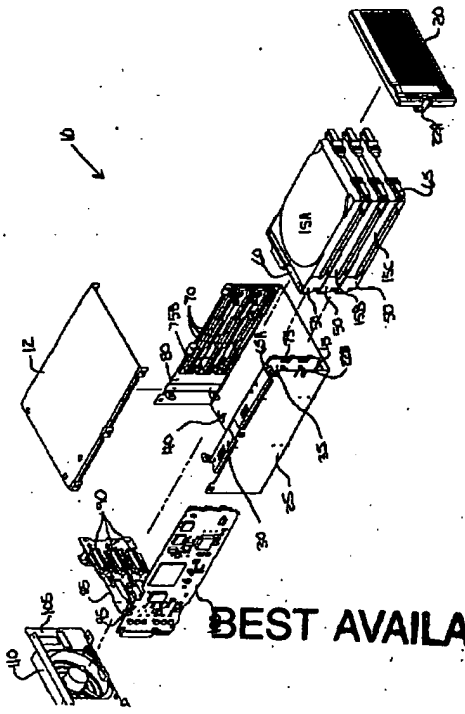
最終頁に続く

(54)【発明の名称】　高集積度コントローラを有する高密度ＲＡＩＤサブシステム

(57)【要約】

【課題】　この発明はＲＡＩＤアレイのような大容量記憶装置サブシステムのための方法と装置を提供する。

【解決手段】　この発明は、第１および第２の空洞を画定するハウジングを含み、前記第１空洞は、ＲＡＩＤコントローラのようなアレイコントローラを収容する。第２空洞は、３．５インチ規格に適合する実質的に通常のＩＤＥドライブを複数個収納する。このアレイは、ハウジングにより限定される極端に小さな空間内で、アレイコントローラとドライブの冷却を最大にするように構成される。



【特許請求の範囲】
【請求項1】　ホストシステムへのＳＣＳＩインターフェイスと内蔵大容量記憶装置へのＩＤＥインターフェイスを含む大容量記憶コントローラを収容する縦長の第1空洞と、
前記第1空洞の縦方向の軸が第2空洞の縦方向の軸に平行になるように、前記ホストシステムと通信可能な複数のＩＤＥドライブを収容する前記第2空洞と、
そこを貫通するベントを有し、前記第2空洞内に収容されたＩＤＥドライブの各々に接続するバックプレーンと、
前記第1空洞と前記第2空洞を通る充分な空気流を供給するために、前記バックプレーンとケースにより形成されるプレナムチャンバーを含む、
大容量記憶アレイサブシステム。
【発明の詳細な説明】
【０００１】
【発明の属する技術分野】本発明はディスクドライブに関し、特にＲＡＩＤアレイサブシステムおよびコントローラに関する。
【０００２】
【従来の技術】ハードディスク記憶装置は、全てのパーソナルコンピュータおよびサーバ、また同様に多くの他の関連するタイプのシステムに普及するようになった。多くの例において、そうした記憶装置は、少なくとも複数のバックアップとバックアップとの間の時間のためのミッションクリテイカル（ｍｉｓｓｉｏｎ－ｃｒｉｔｉｃａｌ）情報の保管のみを行う。結果としてこれらの記憶装置は信頼性が高く、極度に高度なデータ保全を維持しなければならない。
【０００３】データ破壊を保証するために、多くのタイプの記憶装置サブシステムが開発されてきたが、その中にはミラードドライブ（ｍｉｒｒｏｒｅｄ　ｄｒｉｖｅ）、フェイルオーバ（ｆａｉｌｏｖｅｒ）システム、多重冗長ドライブサブシステムがある。その高い信頼性の故に特によく注目されてきた多重冗長サブシステムの一形式は、「安価なドライブの冗長アレイ」、すなわちＲＡＩＤサブシステムである。
【０００４】一般にＲＡＩＤサブシステムはサーバその他のコンピュータシステム内に組み込まれてきた。一般に、ＲＡＩＤサブシステムは二つまたはそれ以上のディスクドライブ（一般的に同一容量で、しばしば同一タイプ）を含み、そして少なくともＲＡＩＤ構成のいくつかの形式において、各ドライブがサブシステム上に記憶されるデータの最初の部分のための一次記憶装置として働くように構成され、またこのデータの第二の部分のためのバックアップ記憶装置として働くように構成される。ＲＡＩＤシステムのための種々なバックアップスキームが開発されてきたが、その中にはＲＡＩＤ０、ＲＡＩＤ１、ＲＡＩＤ５がある。ＲＡＩＤ０においてはデータの冗長性が全く与えられず、ＲＡＩＤアレイの容量は単純に個々のドライブの容量の合計である。ＲＡＩＤ１においては、ミラードドライブ（ｍｉｒｒｏｒｅｄ　ｄｒｉｖｅ）によく似た結合したしたドライブによってバックアップされる。ＲＡＩＤ１はほとんどの場合偶数のドライブで構成される。これに対してＲＡＩＤ５は、典型的に最少三つのドライブから始まる種々な数のドライブにより構成される（二つのドライブは単にＲＡＩＤ１に程度が下がるだけである）。５ディスクＲＡＩＤ５サブシステムにとって、各ドライブはその容量の８０％が一次記憶装置として働き、またその容量の２０％が二次メモリとして働く。結果として、そうしたアレイの記憶容量はドライブの容量の合計の８０％である。
【０００５】一般に先行技術のＲＡＩＤサブシステムは、サーバの外部にあった。これはとりわけ、スペースと信頼性の問題を生じた。通常のサイズのＰＣケースは一般的にディスク記憶装置のために非常に限られた数のベイしか持たず、また通常のＲＡＩＤは利用可能なスペースに適合するにはあまりにも大きい。これはすでに混雑した領域内に特別な設置スペースを必要とし、また、ＲＡＩＤデバイスへのサーバまたは他のＰＣを接続するための外部ケーブルのためのスペースも必要とする。より一般的な外部装置の故障の原因の一つはケーブル故障であり、これは、しばしば人間がケーブルに衝突したり不適切に切断したりすることによる。
【０００６】いくつかの例において、例えばＨＰネットサーバラインのいくつかのモデルにおいて、記憶装置のために特別のベイを供給する特大のケースが供給されてきた。例えば、ネットサーバＬＭ製品は、このサーバの拡張スロット内に挿入されるＲＡＩＤコントローラ付きの二倍幅ケースと、３．５インチ規格に合致するドライブのための八つのベイのスタックを有する。しかしながらこの解決方法は明らかに特定ベンダーの特定モデルのサーバを買うことを要しており、こうしてユーザのオプションを制限する。その上、このＲＡＩＤコントローラは、他の装置に使用できる拡張スロットを占有することとなる。先行技術のこれらの制約は、ＲＡＩＤサブシステムを既存のサーバ内に含めるというユーザの希望を、非常に限られたオプションとしてきた。
【０００７】この発明の譲受人は、既存サーバ内にＲＡＩＤサブシステムを含ませる際にエンドユーザが直面する二律背反のいくつかの面を解決することを、これまでに試みてきた。たとえばアイワ／コア（ＡＩＷＡ／ｃｏｒｅ）のマイクロアレイは、５．２５インチ標準高さ規格内に収まるように構成された。これによりこのサブシステムは、大部分の既存のケース内に設置可能になり、先行技術における置き場所（ｆｏｏｔｐｒｉｎｔ）と外部接続の問題を避けることができた。このマイクロアレイ製品は、２．５インチ規格に合致する（最高５台までの）複数のＩＤＥディスクドライブを、このサブシステ

ムへ挿入することを可能にする。このマイクロアレイ製品は、ＲＡＩＤコントローラと、ＩＤＥドライブをこのＲＡＩＤコントローラにインタフェースしホストシステムへの外部ＳＣＳＩインターフェースを供給するための関連電子製品とをその５．２５インチ規格内に有する。

【０００８】このマイクロアレイ製品は既存の先行技術に勝る多くの利点を提供するが、いくつかの欠点も有する。一つの重大な欠点はそれが高価な２．５インチディスクドライブを必要とすることであり、この２．５インチディスクドライブは、一般的に３．５インチ規格に合致するドライブよりもはるかに容量が少なく信頼性が低いもので、一方同時にかなりコストが高くなる。これらの制限の故に、２．５インチドライブは典型的にラップトップアプリケーションのみにおいて市場を見出したが、一方大部分のデスクトップアプリケーションは３．５インチドライブを使用してきた。

【０００９】その上、マイクロアレイ製品のＲＡＩＤコントローラは、今日他の装置において利用できるものに比較して、制限されたスループットしか提供されておらず、複雑で従って高価な設計となっている。このコントローラは実質的に従来の考えをインプリメントし、またアレイの中の各ドライブごとに独立のＩ／Ｏチャネルを提供する。これにより大きなスペースを必要とし、２．５インチ規格よりも大きいドライブの使用を妨げている。

【００１０】

【発明が解決しようとする課題】結果として５．２５インチ標準高さのベイの従来のサーバケースに適合でき、同時にそのスペース内に統合されたコントローラを提供し、また低コストで大容量の３．５インチドライブの使用を可能にするＲＡＩＤサブシステムへの需要が存在してきた。

【００１１】

【課題を解決するための手段】この発明は先行技術を大きく改良するＲＡＩＤサブシステムを記述するが、その改良点は、改良した容量、改良したスループット、より高い信頼性、より低いコストであって、また同時に単一の５．２５インチ標準高さのベイの中にフィットするものである。この発明のＲＡＩＤサブシステムは、ＥＩＤＥインタフェイスを使用する複数の３．５インチディスクドライブの使用を含み、また同時に望ましい高速のデータ転送速度を有するホストシステムへのウルトラＳＣＳＩインターフェイスを提供するものである。

【００１２】前記の目的を達成するために、機械的および電気的なインターフェイスの注意深い管理が、アレイ中の個別のドライブとコントローラの間に、またサブシステムとホストの間に必要とされてきたが、これは厳しく制限されたスペース内で所望の性能を得るためである。その上サブシステム内で通風のために利用できる空間が非常に限られているので、注意深い熱管理が必要とされてきた。最後に前記要件が従来のコントローラ設計の使用を実質上不可能にしたので、この発明の一部分として高度に集積したＲＡＩＤコントローラが開発された。この発明のコントローラは追加的な特徴として、この発明のＲＡＩＤサブシステムの機械的な設計の外側の領域に大きな利点を提供する。

【００１３】上記の機械的、電気的、熱的、諸問題に加えて、この発明はエンドユーザによるメンテナンスが容易にできることを目的としており、サブシステム内に統合されたドライブへエンドユーザが容易にアクセス可能とする付加的な要件が加わる。これはエンドユーザがサブシステムのフロントパネルを取り除いて、一つまたはそれ以上のドライブを取り除くことを可能にする。これは、１９９７年９月１６日に出願され、「ディスクドライブラッチ」と称し、この発明と同一の譲受人に譲渡され、本願に参考文献として組込まれた米国特許出願第０８／９３１７６６号に記述された方法である。同時に、各ドライブの動作についての情報に関しエンドユーザは、少なくともこのサブシステムフロントパネルにステータスとアクセスの情報が配信されることを強く希望する。ユーザにそうした情報を提供する最も信頼できる方法は、ＲＡＩＤコントローラが装着されるプリント基板にＬＥＤまたは他のディスプレイデバイスを集積することであるが、そうした設計を行えば、エンドユーザが少なくともそのプリント基板の端に触れるかもしれない。この結果として、アクセス中にエンドユーザがサブシステムの内部に適当な注意をしない場合、大量の静電気放電すなわちＥＳＤから、コントローラ基板を保護しなければならない。

【００１４】前記のようにこの発明のコントローラに対しては、先行技術に一般に見出されない多数の設計上の制約がある。これらの中には、空間の制限があり、それは制御基板のための規格内で利用できる空間が、単純に従来のコントローラの設計の使用を許容しないことである。第二に規格により課される熱の要件が、従来のコントローラの設計を過剰な熱を生ずるものとして受け入れ不可能にしている。第三にコストの要件が、複数のコントローラの使用を望ましくないものにしている。

【００１５】結果として、高度に集積されたＲＡＩＤコントローラが開発され、そこではＳＣＳＩホスト機能により使用される単一のＩ／Ｏチャネルが供給され、またアレイ内に含まれる複数のドライブと共にＤＭＡ機能が供給される。この単一のＩ／Ｏチャネルは、時間多重化されていて、これにより各ドライブが所定の限られた時間にコントローラへアクセスできるようになっており、またインタフェースのＳＣＳＩホスト部分が、同様に所定の限られた時間にコントローラへアクセスできるようになっている。適当なクロック速度を使用することにより、この単一チップコントローラは、それに必要な諸機能の各々に対応すると同時に、必要なＤＭＡ機能を管理

することができる。一つの実施例において、コントローラのエンジンは、在庫品のフィールドプログラムマブルゲートアレイ、すなわちＦＰＧＡで実施できるが、その設計はＡＳＩＣまたは他の類似の装置によっても実施できる。ここではこの発明のコントローラを内蔵ＲＡＩＤサブシステムに使用する例を示したが、内蔵および外部の両方のＲＡＩＤサブシステムへの適用、または全く外部のＲＡＩＤ環境への適用にまでこの設計はおよぶ。

【００１６】その上、この発明のアレイは、アレイ内に維持されるディスクドライブのホットスワッピングを可能にする。ユーザがアクセス可能なドライブ特定スイッチにより、このシステムのファームウエアはドライブをパワーダウン可能である。それからこのドライブは除去されて、新たなドライブが設置される。それからファームウエアは自動的に新ドライブの設置を検知して、電力を再び印加すると共に、データと制御信号を再接続する。この技法は、ダウンタイムやデータの損失なしに遂行できるメンテナンスを可能にし、パワーサージを抑制し、静電気放電からの保護を行う。

【００１７】この発明の、これらおよび他の特徴は、以下の「発明の実施の形態」を添付図面と共に参照することにより、一層良く理解されるであろう。

【００１８】

【発明の実施の形態】全体に図１ないし図８、特に図１と図２を参照すると、この発明のＲＡＩＤサブシステムが、より良く理解できる。後でより良く理解されるように、このサブシステムのトップカバー１２は、図１では除去されているが、図２に明示されている。複数の従来型ＩＤＥコンプライアントディスクドライブ１５Ａ、同１５Ｂ、同１５Ｃ（ＩＤＥはその一般的な範囲にＥＩＤＥとウルトラＤＭＡドライブを含む）は、その各々が受け入れられた３．５インチ規格に従っており、ケース２５内のフロントベゼル（ｂｅｚｅｌ）２０の後ろに装着されている。ケース２５は、ベゼル２０と共働して、ほぼ幅５．２５インチ、高さ３．２５インチとして一般に受け入れられている、５．２５インチ規格の標準の高さに一致する。ラッチ２２Ａは、ベゼル２０と一体に形成され、ケース２５内で受け皿２２Ｂと噛み合わされているが、このラッチ２２Ａは、ケース２５の内側に係合するベゼルの反対端の内側のＬ型ポスト（図示なし）と共働して、ベゼルがラッチから外れて、外へ開いてメンテナンスのために、取りはずされるようにする。規格の長さの制約はもっと緩いが、一般に８インチないし１０インチ程度である。内部の上板３０と内部の側壁３５が、ケース２５に固定されて、３．５インチドライブ１５Ａないし１５Ｃを装着するに適した第１空洞４０を画定する。上板３０と側壁３５はまた、第１空洞４０の左に、長く狭い第２空洞４５を囲むが、これについては後で詳述する。

【００１９】各ドライブ１５Ａないし同１５Ｃは、Ｕ型ドライブブラケット５０（図８と図９に最も良く示されこれに関連して詳細に説明されている）内に装着され、Ｕ型ドライブバスケット５０は一組のレール５５Ａ－Ｂおよびドライブ拡張ボード６０を含んでなる。装着機構６５がレール５５Ａ－Ｂ上に装着され、この機構は１９９７年９月１６日に出願され「ディスクドライブラッチ」と題し本願に参考文献として組み入れられた米国特許出願第０８／９３１７６６号に記述されている。レール５５Ａ－Ｂは、装着プレート７５Ａ－Ｂ（図２と図４に最も良く出ている）に合わせて複数の溝７０内に、滑動可能にはめ込まれており、装着プレート７０Ａ－Ｂは、ケース２５の右側壁８０の内側と内部側壁３５の右面に取り付けられている。

【００２０】ドライブ１５Ａないし同１５Ｃの各々に結合されたドライブ拡張ボード６０の後ろに位置しているのは、バックプレーン８５で、図６と図７に関連して以下に説明する。バックプレーン８５は、各ドライブ拡張ボード上でマッチングコネクタ９０Ａと噛み合わせるために、複数のコネクタ９０（図２と図７に特に示される）を有し、またケース２５の左側の下の空洞４５内に装着されるＲＡＩＤコントローラプリント基板１００を装着するために、コネクタ９５（図６と図７に最も良く出ている）を有する。バックプレーン８５とＲＡＩＤコントローラ基板１００を収納し、ファン１１０を支持するために、後ろカバープレート１０５がケース２５の後ろに取り付けられている。後ろカバープレート１０５は、バックプレーン８５の後ろに間隔をとって、プレナムチャンバー１１５を形成し、これによりケース２５により窮屈に配置されたＲＡＩＤコントローラ基板１００および複数のドライブ１５Ａ－Ｃをファンが効率的に冷却できるようにしている。上記の種々の素子の他の詳細は、他の図面に関連して以下に説明する。

【００２１】続けて全体に図１ないし図８を参照し、また特に図３を参照すると、ディスクドライブ１５Ａ－Ｃの配置とバックプレーン８５に対するそれらの接続がより良く理解できる。ドライブ１５Ａ－Ｃ（ドライブ１５Ａのみが図３に示されている）はラッチ機構６５によりケース２５の中にラッチされ、ラッチ機構６５はドライブ拡張基板６０に取り付けられたコネクタ９０Ａをバックプレーン８５上のコネクタ９０と噛み合い接続させる。ドライブ拡張ボードがドライブ１５Ａの後ろにいくらか間隔をとっておかれ、これにより特に複数のドライブ１５Ａ－Ｃの長さのバラツキを許容し、また気流チェンバも形成している。同様にコネクタ９０と同９０Ａの空間もドライブ拡張ボード６０とバックプレーン８５の間に気流チェンバ１５０を形成する。ドライブ１５Ａが、フレキシブルリボンケーブル６０Ａにより、ドライブ拡張ボードに接続されているのがここに見えるが、図９には一層よく見える。リボンケーブル６０Ａは、ドライブ１５Ａに含まれるＩＤＥコネクタに接続し、また異

なったタイプのドライブ上のコネクタの位置の僅かな変動を許容する。

【００２２】バックプレーン８５は、（左で）装着ブラケットの上部および下部ペア１５５と、もう一つのペア１６０により、ケース２５に取り付けられる。例示的な実施例において、内側側壁３５と一体に形成された装着ブラケット１５５は、二重に曲げられている。装着ブラケット１６０は、側壁８０へ取り付けられている。多くの場合必要ないが、装着ブラケット１５５内のこの二重の曲げにより提供される弾性は、挿入および除去の処理によりドライブとバックプレーンに加えられる歪みの力を吸収するのを補助する。更に、装着ブラケットとバックプレーンの弾性は、リボンケーブル６０Ａと共に、ファンまたは他のドライブまたはシステム内の他のどこからかにより加えられる全ての振動から、ドライブを分離するのを補助すると考えられる。この組み合わせは、システムの信頼性を増し、ドライブの寿命を延ばすのを補助すると考えられる。少なくとも、いくつかの場合において、バックプレーン８５とドライブ拡張ボード６０の柔軟性は、リボンケーブル６０Ａと共に、適切な柔軟性を与えると共に分離を行うのに充分である。

【００２３】プレナムチャンバー１１５もまた図３から理解され、ファン１１０の正面の減圧空間に見ることができる。プレナムチャンバー１１５は、空洞４０を通じてドライブ１５Ａ－Ｃの周りに引き込まれ、空洞１５０内に集められた空気と共に、空洞４５を通じてＲＡＩＤコントローラ基板１００を通過して引き込まれた空気を集める。バックプレーン８５と後部カバープレート１０５の間の間隔は、ファン１１０がＲＡＩＤアレイを通じて空気を引き込み、許容可能な温度範囲内にアレイを維持する効率を最適化するよう必要なだけ調節できる。

【００２４】製造を容易にするために、ＲＡＩＤコントローラ基板は、空洞４５内に滑動可能に装着される。２組のガイド１６５は、上部壁３０に下向きに穴をあけて実質的にスロットを形成することにより、単一的に形成可能であり、ケース２５の底に形成された類似のスロット（図示なし）と組み合わせて、基板１００の上縁を空洞４５内の中心に位置させる。同様のガイド１７０もまた壁３０の正面に設けられる。

【００２５】図４を参照すると、ドライブ１５Ａ－Ｃをスタックする配列が、空洞４０と４５を通る空気流と同様により良く理解できる。図３では、上部カバーが示されていない。ＲＡＩＤコントローラ基板１００は、空洞４５の中央に配置され、基板のいずれの側からも空気流が通れる。その上、装着ブロック７５Ａ－Ｂとレール７０の間の間隙が、空洞４０内で、ドライブ１５Ａ－Ｃのいずれの側にも空気を通過させる。空洞４０と４５を通る空気流に一致するように、ファン１１０とプレナムチャンバーのサイズを適当に定めることにより、これらのドライブとＲＡＩＤコントローラ基板は充分な冷却がなされて、長期連続運転を可能にする。より薄いドライブを使用して、図１１に関連して説明するＲＡＩＤコントローラに相応な変更を行えば、ドライブを追加し得る。

【００２６】更に、リーフスプリング１７５が、空洞４５の正面に配置され、それにより基板１００を正しい位置へ付勢すると共に、アレイのメンテナンス中にユーザにより基板に印加されうる静電気電荷を放電すべく、基板１００のアースも行う。コンピュータシステム内の大部分のサブシステムと異なり、ＲＡＩＤコントローラ１００の前縁は、ベゼルをはずすだけで、コンピュータシステムのフロントパネルからユーザによりアクセス可能となる。結果として、ＥＳＤ用の接地に適当なパスが、基板１００の１方側の少なくとも一部分を導電物質でメッキして、そのメッキをリーフスプリング１６５を通じてケースに接続することにより提供される。このリーフスプリングは、一般的に、銅または他の適当なスプリング材料で形成し得る。少なくともいくつかの場合に、そうした他の材料の銅メッキが望ましい。基板１００のメッキは図１０Ａに最も良く示され、参照番号２８５でメッキが識別される。

【００２７】更に図４に示されるのは、各ドライブの１組のＬＥＤ１８Ａ－Ｂと、各ドライブの押しボタン１８５である。ＬＥＤ１８５Ａは典型的に、関連するドライブのステータスを示し、異なった動作状態を示すのに異なった色を用いる多色ＬＥＤである。ＬＥＤ１８０Ｂは一般的に、関連するドライブの活動を示す。押しボタン１８５は、ＲＡＩＤコントローラに、ユーザによる関連ドライブの切断指示信号を与える。押しボタン１８５を押すことにより、ＲＡＩＤコントローラは関連ドライブへの電力と信号パスを切断して、残りのアレイを動作し続けながら、そのドライブを安全に取りはずしできるようにする。ドライブが一旦アレイから電気的に切断されると、ラッチ６５により、このドライブが物理的に取りはずし可能状態となる。このドライブまたは他の同等なドライブが、それからラッチ６５を締めることによりアレイへ戻される。例示的な実施例において、バックプレーンコネクタ９０へドライブコネクタ９０Ａを挿入すると、アレイによりドライブの付加が検知される。しかしながら、いくつかの実施例においては、再び関連の押しボタンを押すことにより、新しいドライブの追加をアレイが検知するようにしてもよい。

【００２８】次に図５を参照すると、このサブシステムの後部は上面半横向き透視図に見られ、冷却ファン１１０とホストを接続するための外部コネクタを、より良く理解できるようになっている。図３に関連して説明したように、冷却ファン１０はバックプレーンの中央の背後に配置され、ファンからの受け入れがたい乱気流を充分に避け得るように間隔をあけてあり、ドライブとプリント基板を通る空気流の量を増加させ、従ってファンの効果を最大化している。ファン１１０の左に、９ピンＤシ

ェルコネクタ２００が配置され、これはこの発明の譲り受け人により提供されるアレイビュー（Ａｒｒａｙ　Ｖｉｅｗ）製品またはサブシステムのステータスをモニタするのに適した他の製品のようなモニタ装置に接続されて使用される。Ｄシェルコネクタ２００の下には、通常のパワーコネクタ２０５がある。Ｄ型コネクタ２００とパワーコネクタ２０５は、例示的な実施例において、バックプレーン８５に接続され、後部カバープレート１０５の開口を通じて延びている。ファンの右側に、シングルエンド超ワイドＳＣＳＩ標準に合致した高密度コネクタ２１０があり、これと共に、ユニットのＩＤの設定、種々の診断および他の通常の機能の遂行のためにふさわしいジャンパーブロック２１５がある。コネクタ２１０と２１５は、ＲＡＩＤコントローラ基板１００に取り付けられ、後部カバープレート１０５の開口を通して延びている。ＳＣＳＩコネクタ２１０は、ホストシステムへのインターフェイスを提供し、またこのサブシステム全体が、ホストシステム内のホストアダプタに対して単一のＳＣＳＩ装置に見える。他の実施例において、このサブシステムは、ディファレンシャルＳＣＳＩ、ワイドＳＣＳＩ、または他のインターフェイスなどの異なったコネクタの異なったインターフェイス規格に準拠するようにして良い。

【００２９】次に図６を参照すると、後ろカバープレート（ファンを含む）を除去した状態のアレイサブシステムの後立面図が示され、それによりバックプレーン８５のレイアウトが詳細に示されている。また図７を参照すると、前立面図にバックプレーン８５のレイアウトを示す。特に図６を参照すると、コネクタ２００と２０５がバックプレーン８５と一体に示され、デュアルコネクタ９５がバックプレーン８５をＲＡＩＤコントローラ基板１００に接続する方法も示されている。その上、多様なベントすなわちカットアウト２２５が、バックプレーン８５の周辺と内部の両方に存在して、プレナム（ｐｌｅｎｕｍ）チャンバー１１５への空気流を改善することが分かる。バックプレーンは孔２３０を貫通する４つのネジ（図示なし）により保持され、装着ブラケットの噛み合うペア１５５と１６０へ装着される。またバックプレーン８５の上にアラーム２３５があり、これはアレイのパフォーマンスをモニタする多様なセンサからの信号に応答する。例えば、これは、１つまたはそれ以上のドライブ温度センサ２４０、ファンセンサ２４５などを含み、実施例では、図７に示すように、バックプレーンの正面に装着される。更に、バックプレーンの正面に示されるコネクタ９０は、高サイクル低挿入力コネクタで、通常のＩＤＥバスとパワーの両方を、関連のドライブに供給する。それからドライブエクステンション６０はドライブへ適当な機械的インターフェイスを供給するが、これには通常のＩＤＥコネクタと通常のパワーコネクタによる。コネクタ９０へ接続するドライブ１５Ａ－Ｃの特有の順序は重要でないが、本例示的な実施例では、トップコネクタに結合されるドライブは、ドライブ０と指定され、ミドルコネクタはドライブ１と指定され、ボトムコネクタはドライブ２と指定される。

【００３０】次に図８と図９を参照すると、単一のドライブ１５Ａをドライブブラケット５０にはめ込む方法がより良く理解できる。図１と図２に関連して前記したように、ドライブブラケット５０は一組のレール５５Ａ－Ｂと共にドライブ拡張ボード６０を有してなる。ドライブ１５は普通の機械ねじによりブラケット５０に取り付けられ、またケーブル６０Ａおよびバックプレーンコネクタ９０と共に通常のＡｍｐｈｅｎｏｌパワーコネクタ６０Ｂを通じてブラケットへ電気的に接続されている。またラッチ機構６５が示される。

【００３１】次に図１０と図１１を参照すると、ＲＡＩＤコントローラ基板１００が見られる。ＲＡＩＤコントローラ基板１００は単一の両面プリント回路基板で、それは図１２からより良く理解される。図１０Ａに示される側はアウトボード（ｏｕｔｂｏａｒｄ）側であるが、こちらから見るとコネクタ２１０と同２１５は左端に見られる。ＲＡＩＤコントローラはＲＡＩＤエンジン集積回路２６０（これはフィールドプログラムマブルゲートアレイ、ＡＳＩＣ、または他の適当な実装であってもよい）を含み、これにより必要な待ち行列とＤＭＡの機能を遂行する。ＲＡＩＤエンジン２６０は、キャシュメモリ２６５（図１１）、ＲＡＩＤコントローラの動作を管理するためのＲＩＳＣ　ＣＰＵ２７０、それに結合したＣＰＵメモリ２７５（両方とも図１１）、ホストインタフェースを管理するためのＳＣＳＩプロセッサ２８０（図１０）と通信する。ＬＥＤ１８０Ａ－Ｂと押しボタン１８５が、前縁でＲＡＩＤコントローラボードに接続されるのが見え（図１１）、一方ボードの反対側に、導体ＥＳＤメッキ２８５（図３に関連して一般に議論される）が見える。本願に示されるＲＡＩＤコントローラボード１００の例示的な実施例は、バックプレーンをボード１００に接続できるようにするために一組のコネクタ２７５を含む。時間／日付チップ２９０と共に、図１２に関連して記述した通常の機能を遂行する種々の他のセンサと論理回路も供給され得る。メッキ２８５の配列から見て、特に図１０から分かることは、この発明のサブシステムのメンテナンスを行うユーザが、メンテナンス中に、静電気を放電して伝達し、結果としてＲＡＩＤコントローラを損傷するのを実質的に防止するようになっていることであり、それはメッキ２８５が上記のように接地面に直接接続されているからである。

【００３２】次に図１２および図１３ないし図１６を参照すると、この発明の電気的動作がよりよく理解される。一般に、ＲＡＩＤサブシステムはホストから見ると、通常のＳＣＳＩコマンドに外部的に適合する単一ボリュームに見えるが、しかし内部的には完全にＲＡＩＤ

アレイとして動作する。このＲＡＩＤアレイの動作はＲＡＩＤコントローラにより制御され、一方ＲＡＩＤコントローラは、時分割多重と独立の３２ビットＤＭＡとＣＰＵソフトウエア処理メモリを使用することにより、エンジンのピーク速度において同時非競合アクティビティを可能にする。ＤＭＡすなわちキャシュメモリ２６５は、たとえば１ｘ３６メモリとして構成された４メガバイトであり、１６０ＭＢ／秒程度の帯域幅を有するシングルサイクルページ化ＥＤＯパイプラインを供給する。ＣＰＵメモリ２７５は、例示的な実施例では１ｘ３２で形成された４メガバイトとして構成され、８０ＭＢ／秒の帯域幅を有する２サイクルページ化ＥＤＯパイプラインを提供する。

【００３３】ＣＰＵ２７０は、たとえば４０ＭＨＺで動作するＬＳＩ　ＬＲ３３３１０－４０　３２ビットＲＩＳＣプロセッサであり、埋め込みＲＡＩＤオペレーティングシステムを記憶するＦＬＡＳＨ　ＲＯＭ３００と共働する。アーキテクチャの中心にＲＡＩＤ集積回路２６０があり、これはたとえばＡｌｔｅｒａフィールドプログラマブルゲートアレイ（ＦＰＧＡ）またはその均等物またはＡＳＩＣとして構成され、これは各ＤＭＡ　Ｉ／Ｏチャネルのためにコマンド待ち行列の提供、種々のデータＩ／Ｏ待ち行列の管理、それに関連するキーバス上のバスアクティビティの管理を行い、システムの周辺機能をサポートする。ＦＰＧＡ２６０に関連する５つの主要なバスは次の通りである：４０ＭＢ／秒の１６ビットＳＣＳＩプロセッサバス３０５（典型的にウルトラＳＣＳＩ動作のために構成されているが他のＳＣＳＩプロトコルをサポートできる）；３．３３ＭＢ／秒の８ビットＳＣＳＩチップパイプライン化Ｉ／Ｏバス３１０；１６ＭＢ／秒の１６ビットＩＤＥドライブバス３１５；１６０ＭＢ／秒の３６ビットディスクキャッシュメモリ（ＤＣＭ）バス３２０；および８０ＭＢ／秒の３２ビットＣＰＵバス３２５である。ＦＰＧＡ２６０は、５つのバス全てが並列に動作できるように構成され、ＳＣＳＩプロセッサバス３０５、ＩＤバス３１５，およびＤＣＭバス３２５によりＲＡＩＤエンジン２６０へのアクセスを多重化するのに充分な速度で動作するＲＡＩＤエンジン２６０を有し、また定義されたサイクル内でＳＣＳＩプロセッサバス３０５とＩＤＥバス３１５の各々にワンタイムスロットをＤＣＭバス３２０のためにツータイムスロットを割り当てることにより、ＩＤＥバス３１５と、ＤＣＭバス３２０を有する。例示的な実施例において、１サイクル全体で１００ｎｓ程度であり、４つのタイムスロットは各々２５ｎｓを割り当てられる。ＲＡＩＤエンジン２６０のパフォーマンスの故に、このサブシステムの正味スループットは、主として以下の４つの要素に依存する：ＩＤＥドライブのパフォーマンス、一体型オペレーティングシステム内のＲＡＩＤ機能オーバーヘッド、ユーザのホストアダプタのパフォーマンス、ユーザホストアプリケーションのドライバオーバーヘッドである。

【００３４】更に図１２を参照すると、システムの動作は実質的に次の通りである：電源投入すると、フラッシュＲＯＭ３００からＲＩＳＣプロセッサ２７０に関連するＣＰＵメモリ２７５へ、オペレーティングシステムをロードすることにより、システムは安定状態になる。初期化の後、ある時点で、読み取りまたは書き込みの要求を、ホストシステム３４０からホストＳＣＳＩバス３５０で受け取り、これが適当でアレイば、終端ブロック３５５により終結される。この要求はそれからＳＣＳＩプロセッサ２８０により処理され、ＳＣＳＩプロセッサ２８０は適当な信号をＲＡＩＤエンジン２６０へＩ／Ｏバス３１０上で送り、適当な確認信号を送り返される。それからこのデータは、ホストシステムによりＳＣＳＩＤＭＡバス３０５上で利用可能にされる。この時点で、ディスクキャッシュメモリ２６５は空である。この要求が情報を書き込むことでアレイば、ＣＰＵ２７０はＲＡＩＤエンジン２６０に、このデータをＤＣＭ２６５へ渡すように命令し、ＤＣＭ２６５でキャッシュ中に保持される。その後、バックグラウンド処理の間に、バス３２０上でＲＡＩＤエンジン２６０により最初にデータをアクセスすることにより、（使用されるＲＡＩＤ　ｓｔｒｉｐｉｎｇに従って、）適当にディスクに割り当てられ、ＩＤＥバス上で、ＩＳＯディスクバッファ３６０Ａ－Ｃへ書き出される。それからこのデータは、特定のディスク１５Ａ－Ｃに書き出される。本願に記述する例示的なシステムにおいては、バス３２０が１０本のアドレス線と３６本のデータ線を含むのが適当である。同様に、バス３２５は、３２本のアドレス線と３２本のデータ線を有している。この処理は、ＲＡＩＤエンジンからＳＣＳＩプロセッサ２８０へ供給され、それからホスト３４０へ供給される確認信号で終了する。これらの種々の事象のタイミングは、図１３ないし図１６および図１７に関連して一層詳細に説明する。図１３において、ＤＰＩとＤＰＯはそれぞれ「入力データパス」と「出力データパス」を表す。

【００３５】読み取り動作においては、この処理は実質上類似しているがいくらか逆になっている。この処理はＳＣＳＩインタフェースをアクテイブにし、一般的には開始の時に行われるようにする。それからホストは確認／肯定応答信号を送り設定を実行し、これに続いてＰＩＯバス３１０上で特定データの要求を送る。それからこの要求はＲＡＩＤエンジン２６０により検出され、ＲＡＩＤエンジン２６０はそれをドライブ１５へ渡す。このデータはこれらのドライブからＲＡＩＤエンジン２６０へ返され、ここからそれは一時的な記憶のためにディスクキャッシュメモリ２６５へ渡される。適当な時間にＣＰＵ２７０は、このデータがバス３２０を介してＤＣＭ２６５から読み出されて、エンジン２６０を通じてＥＳ

ＥＩプロセッサ２８０へデータバス３０５上で渡されるようにする。それからこのデータはＥＳＥＩプロセッサ２８０からホストへバス３５０上で渡される。

【００３６】ＲＡＩＤエンジンはそのデータ管理機能の他に、多数の周辺機器補修機能を管理する。これらの中には温度超過検出器２４０とファンエラー検出器２４５が含まれ、アラーム２３５で（適当な時に）警告信号を生成する。時間／日付クロック２９０もまたモニタされ、その電力は、システムがオフであるときに電力は電池または他の電源３６５により供給される。ハードウエア検出線３７０はステータスレジスタ３７５によりモニタされる。ドライブのための電力サージ制御はバッファ３８０でモニタされる。このサブシステムの監視はまたｄｕａｒｔ３８５上でも供給される。典型的な監視は監視と保守の両方のためにＲＳ２３２リンク３９０上で遂行される。

【００３７】次に図１３ないし図１６および図１７を参照するとＲＡＩＤエンジン２６０の動作の詳細が一層良く理解され、その中にはＲＡＩＤエンジンにより種々なバス上の信号が多重化されるタイミングが含まれる。上記の図と同様に、図１２から同じ部品は同じ参照番号を割り当てられている。

【００３８】前と同様に、ホストシステムの電源が投入されると、ＲＡＩＤサブシステム１０が初期化を行い、フラッシュＲＯＭ３００内に維持されるソフトウエアにより設定されるときにＲＩＳＣ　ＣＰＵ２７０により一連のイネーブリングファクターが生成される。これらのイネーブリングファクターは、ＩＤＥドライブを知られた状態にし、またＳＣＳＩプロセッサ２８０をアクティブでイネーブルされた状態にし、ホストシステム３４０へ通知する。ホストシステムはＳＣＳＩプロセッサからの通知を確認し肯定の応答をする。その上、このイネーブリングファクターは、ＲＡＩＤエンジン２６０を知られた状態にし、また特にＲＡＩＤエンジン２６０の内部にある４０ＭＨｚ　Ｉ／Ｏコマンド待ち行列プロセッサを初期化する。図１５は、図１４の鎖線内の一部分の代案の配列を示す。

【００３９】初期化の後に、ホストシステムは前と同様にドライブ１５Ａ－Ｃに書き込むべきデータを送る。この情報はヘッダ情報とデータを含んでなり、バス３５０上でＳＣＳＩプロセッサ２８０へ供給される。ＳＣＳＩプロセッサ２８０で加工した後に、このヘッダ情報は図１３に破線で示したＲＡＩＤエンジン２６０へ、８ビットプログラマブルＩ／Ｏバス３１０で供給される。このヘッダ情報はＲＡＩＤエンジン２６０へＳＣＳＩ　ＰＩＯ４００を通して送られるが、ＳＣＳＩ　ＰＩＯ４００は入力データパス４０５と出力データパス４１０を有する。入力データパス４０５はマルチプレクサｍｕｘ４１５の片側に接続し、マルチプレクサｍｕｘ４１５は間接的にＩ／Ｏコマンド待ち行列プロセッサ３９０へ入力を供給する。このＩ／Ｏコマンド待ち行列プロセッサ３９０はフレームベースのスクリプトプロセッサであって、ハーフワードのコマンド、行アドレスコマンドおよび列アドレスコマンドを、１６ビットバスを経由してレジスタ３９５へ供給する。レジスタ３９５はまたＲＩＳＣプロセッサ２７０からアドレスを受け取る；プロセッサ２７０はまた、ｍｕｘバッファ３９９を通じて与えられるバス３２５の１０ビットブランチ３９７を経由して、ＤＣＭ２６５へアドレスを供給する。レジスタ３９５の出力は、パイプラインレジスタ４０１とバッファ４０３を通じて、ＤＣＭ２６５へ多重化１０ビットバス（１メガバイトのアドレス空間をアドレスする）を経由して供給され得る。レジスタ３９５の出力は、間接的に、ＲＡＩＤエンジン２６０に関連する他の場所を示す出力データバスを供給するが、これにはｍｕｘ４１５への第２の入力が含まれる。

【００４０】ＳＣＳＩプロセッサ２８０からのデータが供給されるヘッダ情報と並行して、ＳＣＳＩプロセッサ２８０からＳＣＳＩ　ＰＩＯ４００へバス３１０上をＲＡＩＤエンジン２６０へ１６ビットＤＭＡバス３０５経由で供給される。特にデータは４０ＭＨｚで作動する１６ビットツー３２ビットファネル４２０へ与えられるが、これはＲＡＩＤエンジン２６０が３２ビット幅で内部的に動作するからである。データはファネルｍｕｘ４２５の片側へ供給されて、それからＩ／Ｏ待ち行列ｍｕｘ４３０の片側へ供給される。待ち行列ｍｕｘ４３０の出力はフレームベースＩ／Ｏ待ち行列４３５へされるが、フレームベースＩ／Ｏ待ち行列４３５は４０ＭＨｚで作動し２５６ｘ３２で構成されて、１６０ＭＢ／秒のスループットを供給する。Ｉ／Ｏ待ち行列４３５への他の入力は、種々のＩＤＥプリンタ４４０、ＳＣＳＩポインタ４４５、ＤＣＭポインタ４５０を有する。データはＩ／Ｏ待ち行列４３５の出力へ全部クロックされた第１パイプライン出力レジスタ４５５へ供給され、それからＤＣＭｍｕｘ４６０の片側へ供給される。ｍｕｘの出力は第２パイプライン出力レジスタ４６５へ供給され、バッファ４７０を経由して、それからＲＡＩＤエンジン２６０からＤＣＭ２６５へ供給される。

【００４１】ディスク１５Ａ－Ｃの適便な一つに書き込むのに適当になるまで、データはＤＣＭ２６５に記憶され、書き込むディスクは典型的に通常のアルゴリズムに従って、ＲＡＩＤオペレーティングシステムにより決定される。その時に、Ｉ／Ｏコマンド待ち行列３９０が、ディスクドライブへデータを書き込むコマンドを発行する。データはＤＣＭ２６５によりバッファ４７５へ供給され、それからパイプライン入力レジスタ４８０へ供給される。データはそれから第２入力レジスタ４８５と共に、プロセッサ入力ｍｕｘ４９０の片側へ供給される。ドライブを書き込むために、データがレジスタ４８５を通じてｍｕｘ４３０の他の側へ与えられ、それからＩ／

O待ち行列４３５へ与えられる。

【００４２】Ｉ／O待ち行列４３５からのデータはＳＣＳＩ　Ｉ／Oパネル４２０へ供給され、またディスクＩ／Oファネル５００へも供給される。ディスクＩ／Oファネル５００は出力データを、ディスクドライブ１５との通信のために、３２ビットデータ幅から１６ビットデータ幅へ再変換する。ディスクドライブとの通信についてのその他のことは、図１２に関連して説明した通りである。

【００４３】ホストシステム３４０によりＲＡＩＤサブシステムに要求されるもう一つの典型的な動作はＲＡＩＤサブシステムからのデータ検索である。データ検索はホストシステム３４０から開始され、ここでもホストの要求をＰＩＯバス３１０、ＳＣＳＩ　ＰＩＯ４００を介してＲＡＩＤエンジン２６０へ供給し、それからｍｕｘ４１５を通じてＩ／Oコマンドプロセッサ３９０への入力データパス４０５へ供給する。Ｉ／Oコマンドプロセッサ３９０は、それから適当なＲＡＣ／ＣＡＣアドレスをレジスタ３９５経由で供給し、これによりデータが検索されるようにする。

【００４４】ホストシステムにより希望されるデータの適当なアドレスが、ＤＣＭ２６５へ供給される。もしデータがＤＣＭ２６５内に維持されれば、それはレジスタ４８０と同４８５を経由してｍｕｘ４３０へ供給され、それからＩ／O待ち行列４３５へ供給される。Ｉ／O待ち行列４３５からこのデータがＳＣＳＩ　Ｉ／Oファネル４２０へ供給され、ここで出力データが１６ビット幅へ変換される。それからデータはＤＭＡ３０５上をＳＣＳＩプロセッサ２８０へ供給され、最終的にホスト３４０へバス３５０上を出力される。

【００４５】しかしながらホストにより要求されたデータがキャッシュ２６５内に現在維持されていなければ、このデータをディスクに要求しなければならない。この場合は、要求されたデータのアドレスがレジスタ４８０、４８５およびｍｕｘ４３０を経由して、Ｉ／O待ち行列４３５へ供給される。それからＩ／Oの待ち行列４３５の出力が、ディスクＩ／Oファネル５００へ供給されてドライブ１５Ａ－Ｃへ出力される。それからデータは、必要な待ち時間の後にドライブから検索され、その後にドライブから到着するデータが、ディスクファネル５００内で１６ビット幅から３２ビット幅へ変換される。ファネル５００のデータの出力はそれからｍｕｘ４２５の第２の側へ供給され、そこからＩ／O待ち行列４３５へ第２のｍｕｘ４３０を通じて供給される。

【００４６】それからＩ／O待ち行列４３５の出力は、Ｉ／Oファネル４２０を通じて出力データについて上記したのと同じ方法で供給され、その結果このデータは通常の方法でホストシステムへ供給される。フィールドプログラマブルゲートアレイ（ＦＰＧＡ）に基づくインプリメンテーションのために、ＦＰＧＡルートＲＯＭ４９２が供給され、電源投入に際してＦＰＧＡがパーソナル化されるようにする。ＡＳＩＣまたは他のゲートアレイのインプリメンテーションにおいては、そうしたルートＲＯＭは不必要である。同様にＲＡＩＤ　ＯＳがＣＰＵメモリ２７５へ電源投入に際してロードされ、全てのソフトウエア制御がＣＰＵメモリ内に記憶された命令から引き出される。

【００４７】この発明の動作の重要な特徴は、単一のＲＡＩＤエンジン２６０がＳＣＩプロセッサ、ディスク、プロセッサ２７０と通信するのに必要な多重ＤＭＡ待ち行列を管理できることである。この目的はデータをＲＡＩＤエンジン２６０へ供給するキーバスを時間多重化することにより達成される。これは４０ＭＢ／秒以上では作動しない他の装置に比較して、このＩ／O待ち行列が１６０ＭＢ／秒の効果的な速度で動作する故に可能である。これによりＲＡＩＤエンジンはその約４分の１の時間をＳＣＳＩプロセッサとディスクドライブの各々へ割り当てることができ、またその約半分の時間をＤＭＡアドレッシングに割り当てることができる。図１７に示すタイミング図は、この発明の実施例のいくつかの面に重要な時分割多重化を提供するのに必要な位相化アドレッシングを与えるものである。

【００４８】特にＩ／O待ち行列４３５とスクリプトプロセッサの４０ＭＨｚクロックが６００で示され、一方ＲＡＩＤへのアクセスするためのＳＣＳＩ位相が６０５で示されている。ＩＤＥ位相は６１０で示され、一方ＤＣＭ位相は６１５で示される。付加的な機能において、ＩＤＥドライブからアクセス要求されない時にサイクルが起こった場合は、ＤＣＭによる使用のために位相が再割り当てされる。同様にＳＣＩプロセッサがＩ／Oアクセス要求しないサイクルについてはＳＣＩプロセッサに割り当てられる位相がＤＣＭへ再割り当てされる。こうしてこの発明により極度に高いスループットが達成できることが理解されよう。

【００４９】次に図１８を参照すると、この発明のホットスワッピング装置、（アレイの残りの部分が動作し続ける間に一つまたはそれ以上のドライブを除去できる）がより良く理解できる。特に、例えばあるドライブの故障のために、ユーザがドライブ１５Ａ－Ｃの一つを取りはずすことを希望する場合、取りはずすべきドライブに関連する押しボタンスイッチ１８５をユーザが作動させる。これがＣＰＵ２７０に信号を与えて、ソフトウエア制御のもとに作動するＣＰＵ２７０がＦＰＧＡ制御論理回路へ信号を与えて、関連するドライブの１２ボルトと５ボルトの電源７００と７０５をそれぞれ切断させる。その上プロセッサによりデータパス７０１と制御パス７１５が、サブシステムの他の部分から電気的に切断される。この時点でユーザはラッチ６５をやり直し、必要なドライブを除去する準備ができている。

【００５０】ドライブを再設定するには、ユーザはドラ

イブをドライブベイに挿入しラッチ６５をラッチすることにより、処理の機械的部分を逆にたどるだけでよい。双安定ラッチがドライブの再挿入を検知して、新しく挿入されたドライブへの電力と信号の接続を再び加えるように、ＣＰＵ２７０へ信号を与える。この方法で、古いドライブが除去されて、新ドライブが設定され得る。

【００５１】従って、ＲＡＩＤのための新しく新規なサブシステムと、高度に統合されたコントローラが記述されたことが、理解されたであろう。当業者にとっては、これらの教示が与えられれば、開示された発明を組み込んだ多数の代案や均等物が存在することが理解されよう。結果として、この発明は、先行する例示的な実施例により限定されるべきものではなく、むしろ請求項によって限定されるべきである。

【図面の簡単な説明】

【図１】上部カバーを除去したこの発明のＲＡＩＤサブシステムの正面半横向き透視図を示す。

【図２】この発明のサブシステムの種々の構成要素の半横向き分解透視図を示す。

【図３】上部カバーを除去したこの発明のＲＡＩＤサブシステムの上面図を示す。

【図４】前カバーを除去したこの発明のＲＡＩＤサブシステムの前立面図を示す。

【図５】上部カバーを除去したこの発明のサブシステムの後ろ半横向き透視図を示す。

【図６】後ろカバープレートを除去したこの発明のサブシステムの後ろ立面図を示し、特にバックプレーンを示す。

【図７】この発明のバックプレーンのレイアウトを、前立面図で示す。

【図８】単一のドライブと関連の装着ブラケットを、バックプレーンインターフェイス基板と共に、上面半横向き透視図で示す。

【図９】図８の装着ブラケットの透視図を示し、特に、ドライブとブラケットの間のリボンケーブルインターフェイスを示す。

【図１０】ＲＡＩＤコントローラ基板の片側を、レイアウト形式で示す。

【図１１】ＲＡＩＤコントローラ基板の第２の側を、レイアウト形式で示す。

【図１２】ＲＡＩＤエンジンを含むこの発明のＲＡＩＤコントローラを、概略ブロック図形式で示す。

【図１３】図１２のＲＡＩＤエンジンの内部構成を示す。

【図１４】図１２のＲＡＩＤエンジンの内部構成を示す。

【図１５】図１２のＲＡＩＤエンジンの内部構成を示す。

【図１６】図１２のＲＡＩＤエンジンの内部構成を示す。

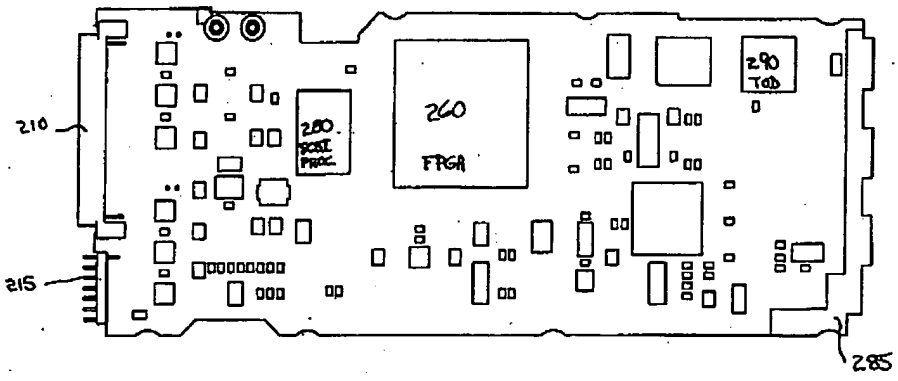
【図１７】図１２に示すＲＡＩＤエンジンの種々の動作のタイミングを示す。

【図１８】ＲＡＩＤアレイのホットスワップ機能を、略図形式で示す。

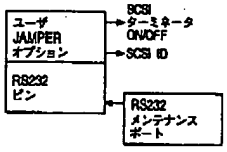
【符号の説明】

１０　ＲＡＩＤサブシステム
１５Ａ、１５Ｂ、１５Ｃ　ＩＤＥドライブ
２５　ケース
４０　第１空洞
４５　第２空洞
８５　バックプレーン
１００　ＲＡＩＤコントローラ基板
１１５　プレナムチャンバー

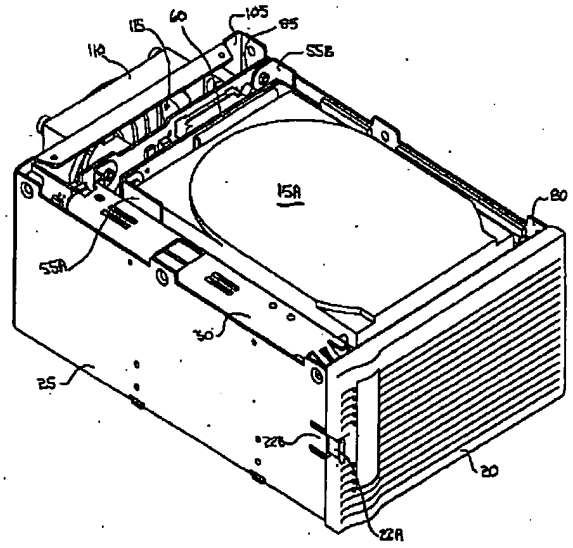
【図１０】



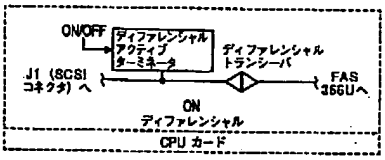
【図１６】



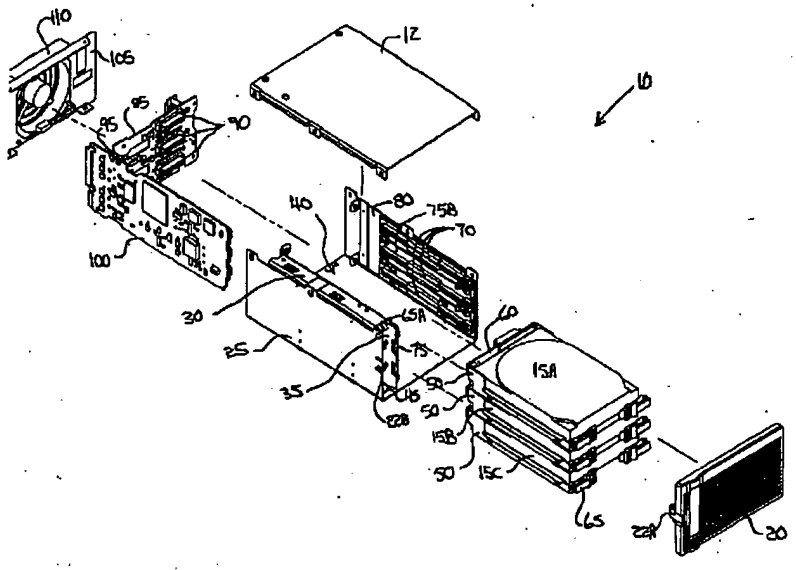
【図1】


110
115
60
105
65
112
55B
15A
80
55A
50
25
22B
20
22A

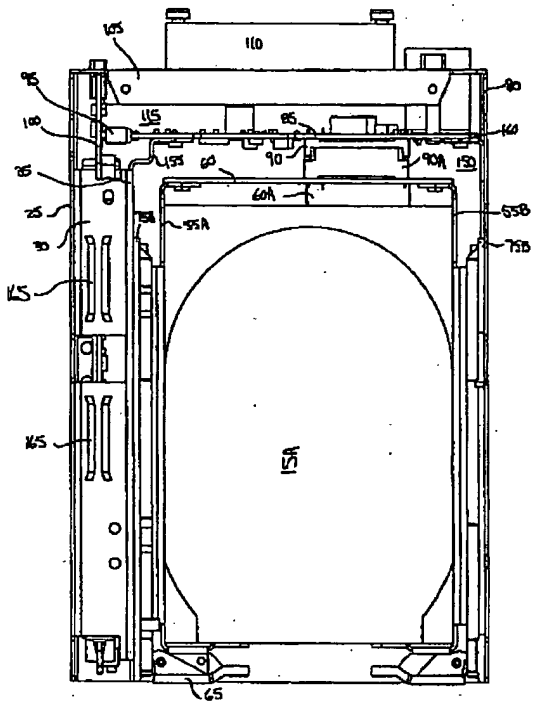
【図15】


ON/OFF
ディファレンシャル
アクティブ
ターミネータ
ディファレンシャル
トランシーバ
J1 (SCSI
コネクタ) へ
FAS
366Uへ
ON
ディファレンシャル
CPU カード

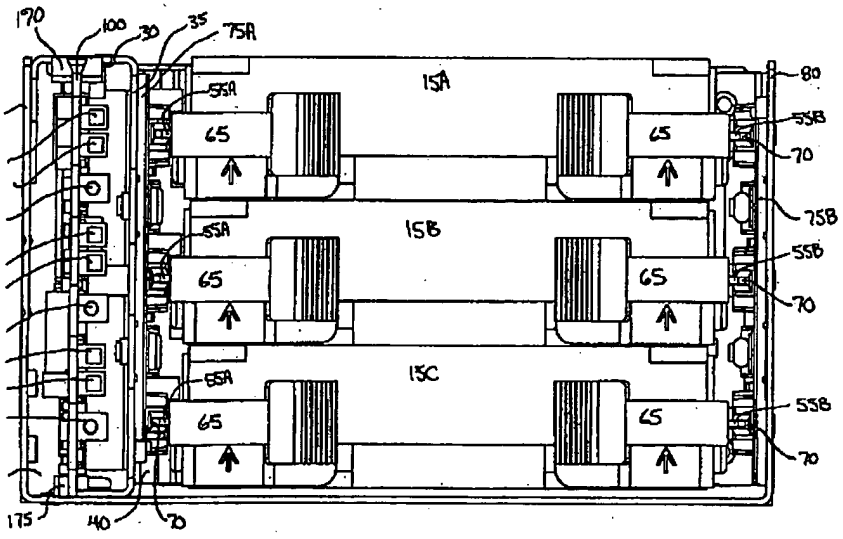
【図2】


110
105
12
10
95
85
90
40
80
75B
70
100
30
15A
60
25
75
35
65
50
22B
15B
50
15C
65
22A
20

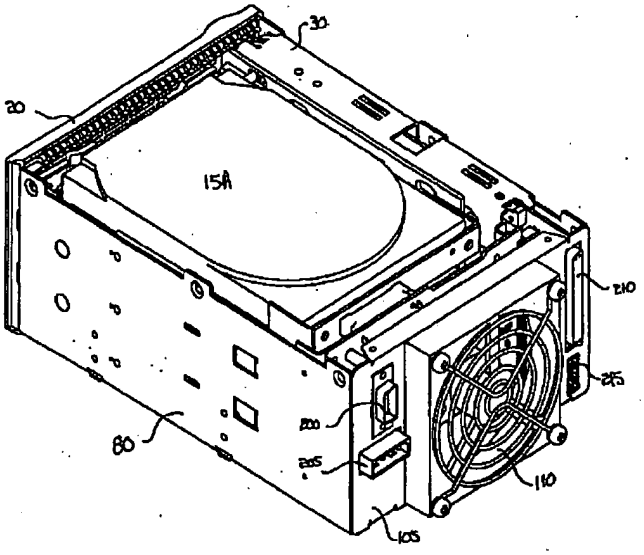
【図3】


110
155
115
85
90
90A
150
60
60A
55B
75B
55A
75A
15S
15A
65

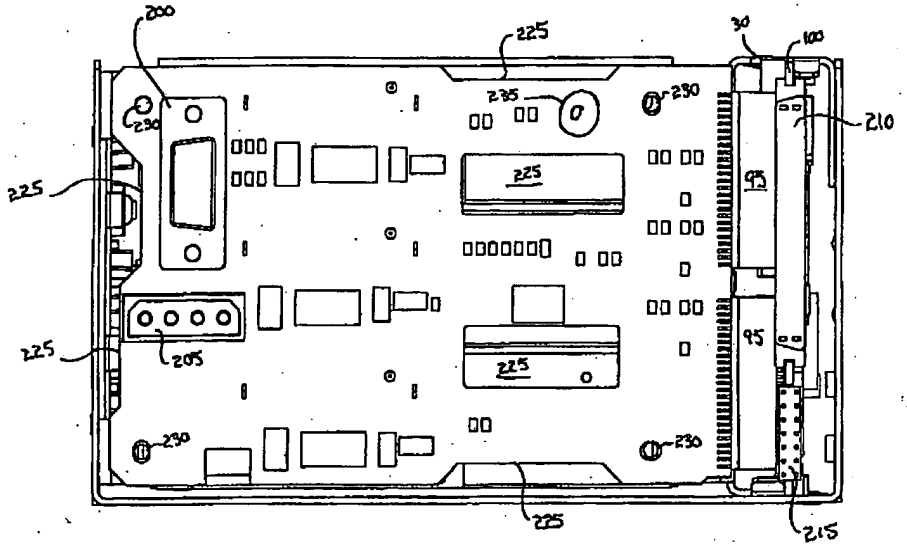
【図4】


170
100
30
35
75A
55A
15A
65
80
55B
70
75B
15B
15C
175
40
70

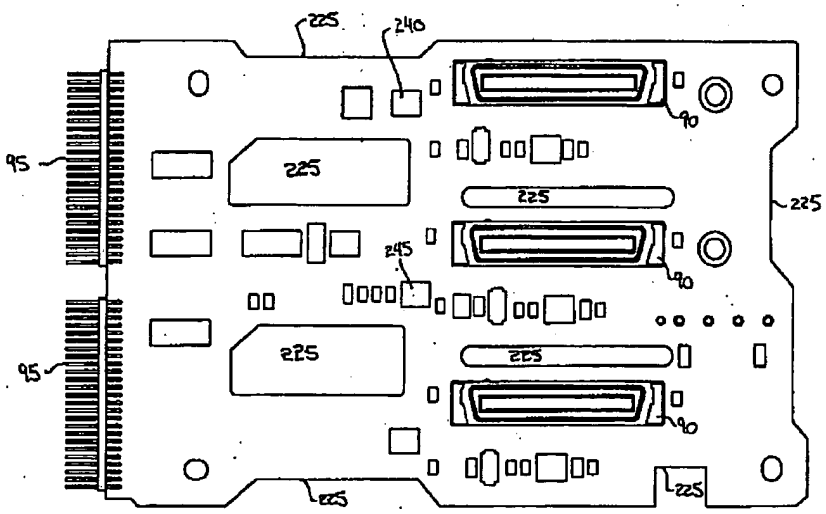
【図5】


30
20
15A
210
215
200
205
80
110
105

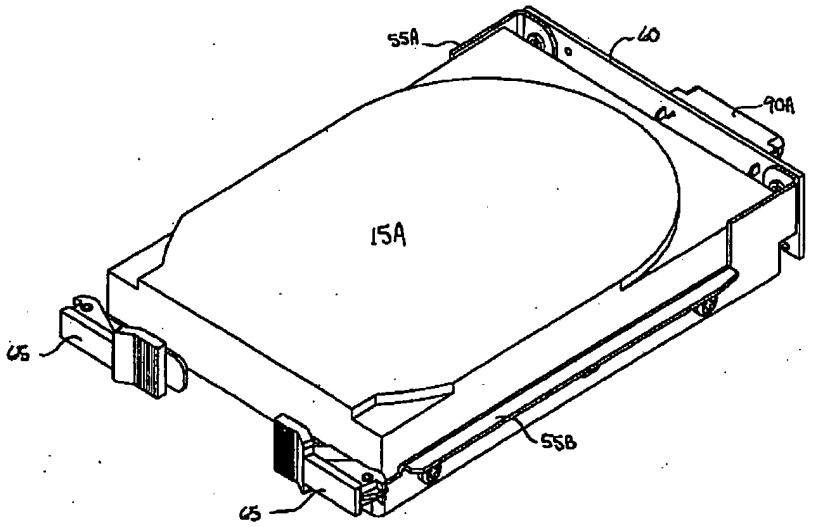
【図6】


200
225
30
100
235
230
210
225
205
95
215

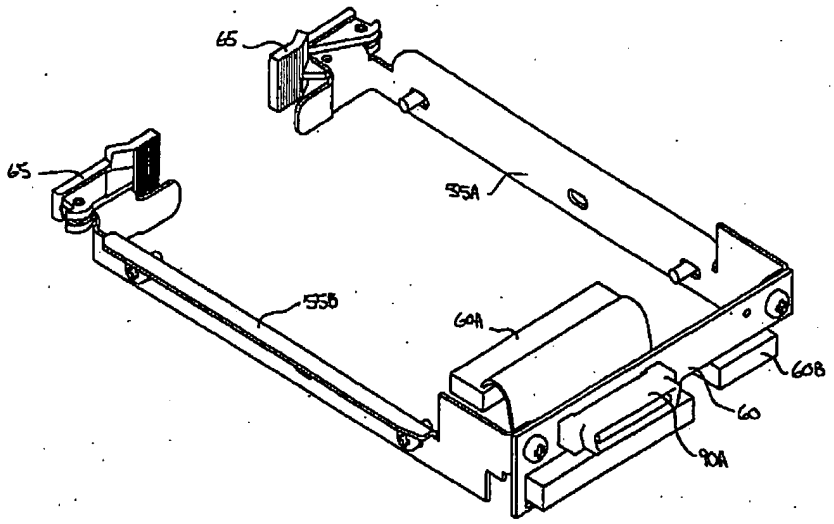
【図7】


225
240
90
95
225
225
245
90
95
225
225
90
225
225

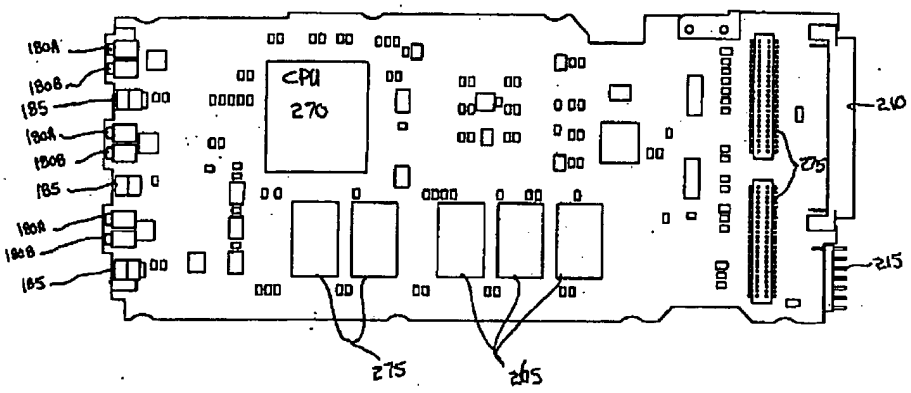
【図8】


55A
60
90A
15A
65
55B
65

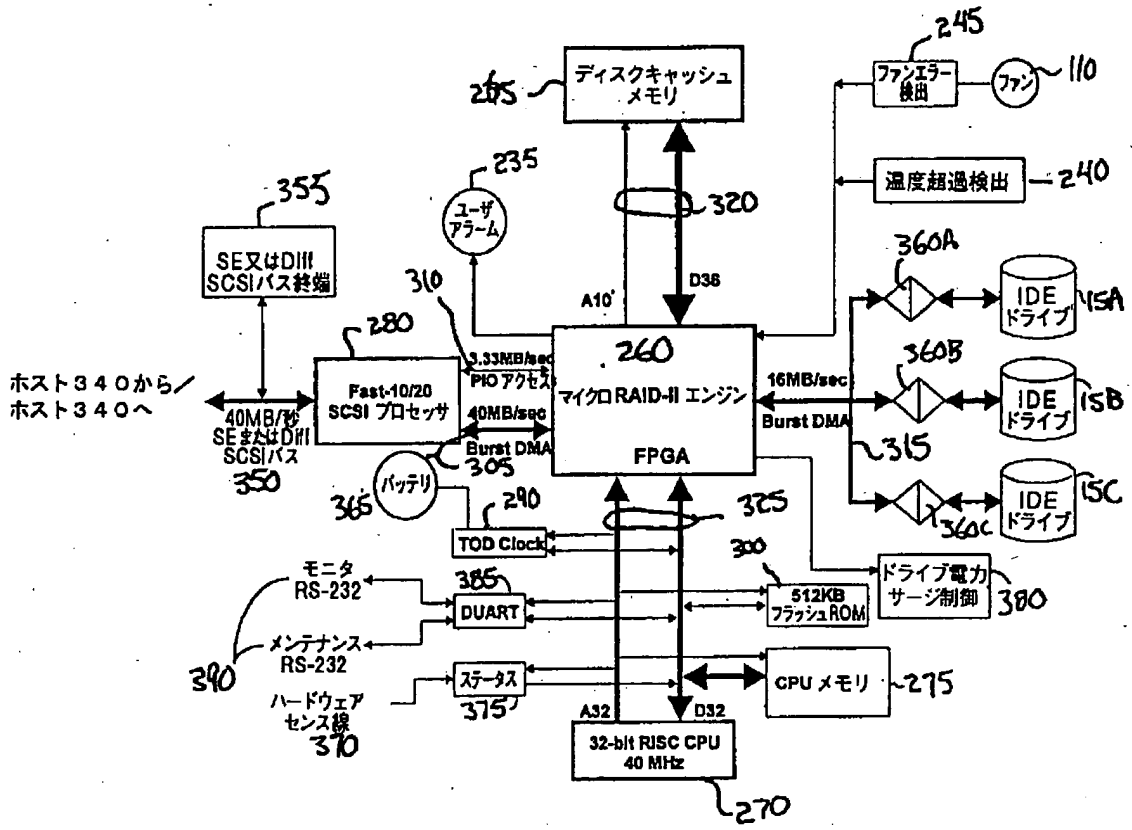
【図9】


65
65
55A
55B
60A
60B
60
70A

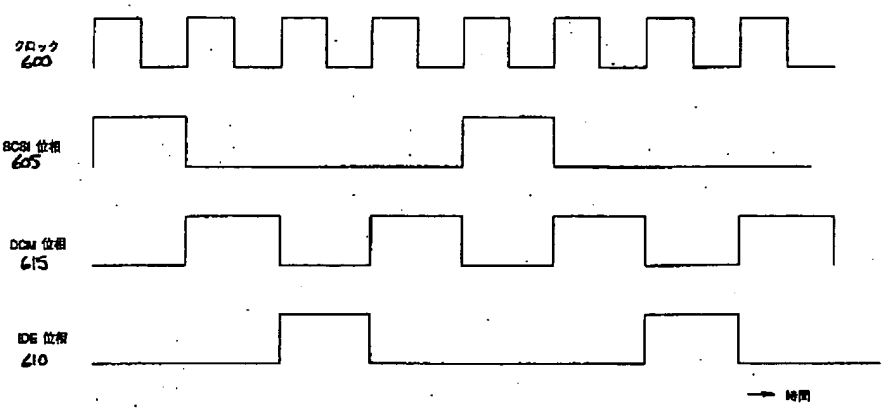
【図11】


180A
180B
185
180A
180B
185
180A
180B
185
CPU
270
210
205
215
275
265

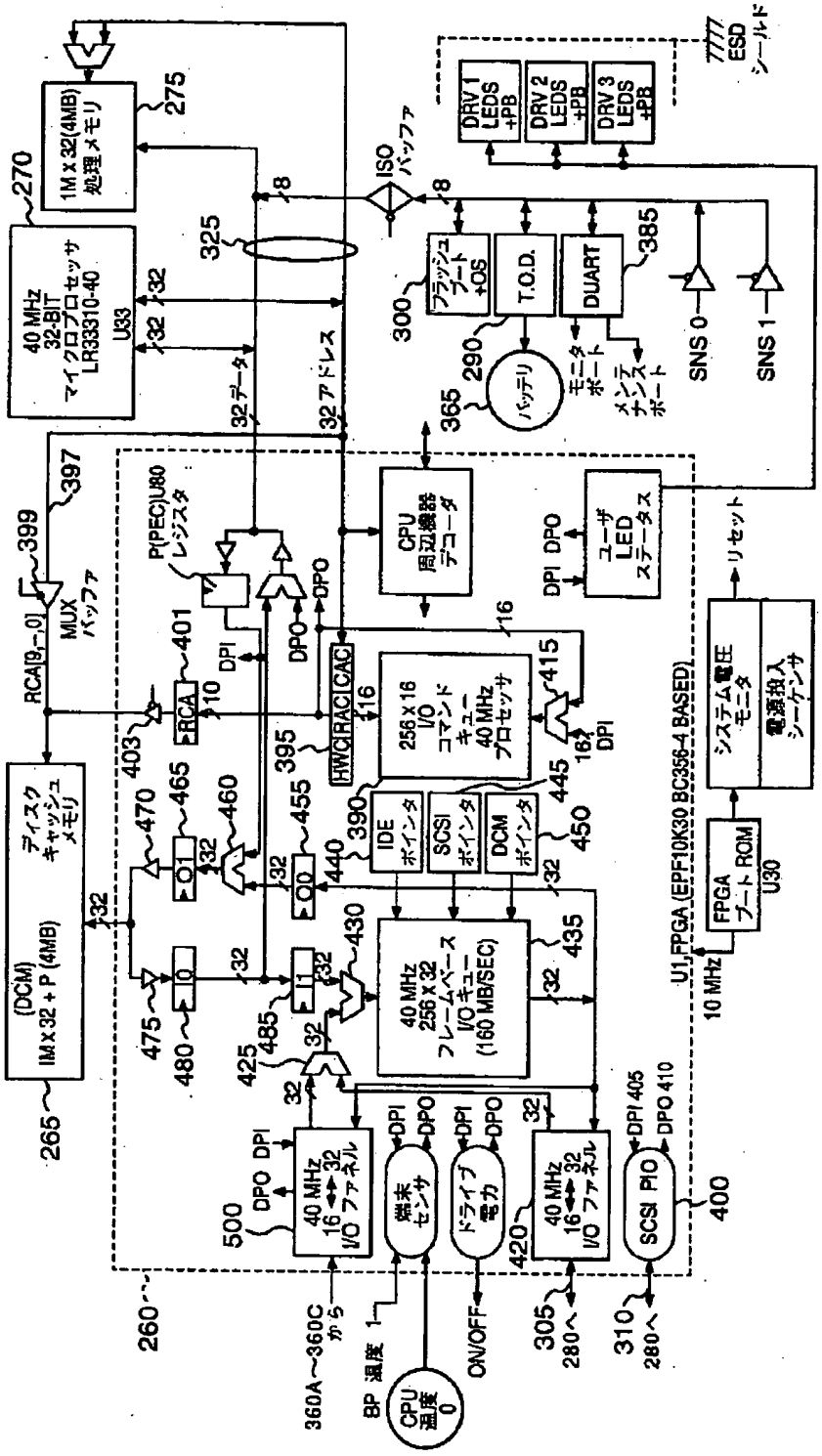
【図12】


ディスクキャッシュ
メモリ
285
245
ファンエラー
検出
ファン
110
温度超過検出
240
320
235
ユーザ
アラーム
355
SE又はDIII
SCSIバス終端
310
A10'
D36
280
260
360A
IDE
ドライブ
15A
ホスト340から／
ホスト340へ
Fast-10/20
SCSI プロセッサ
3.33MB/sec
PIO アクセス
40MB/sec
Burst DMA
マイクロ RAID-II エンジン
FPGA
16MB/sec
Burst DMA
360B
IDE
ドライブ
15B
40MB/秒
SEまたはDIII
SCSIバス
350
305
315
バッテリ
365
290
TOD Clock
325
360C
IDE
ドライブ
15C
300
ドライブ電力
サージ制御
380
モニタ
RS-232
385
DUART
512KB
フラッシュROM
メンテナンス
RS-232
390
ステータス
CPU メモリ
275
ハードウェア
センス線
370
375
A32
D32
32-bit RISC CPU
40 MHz
270

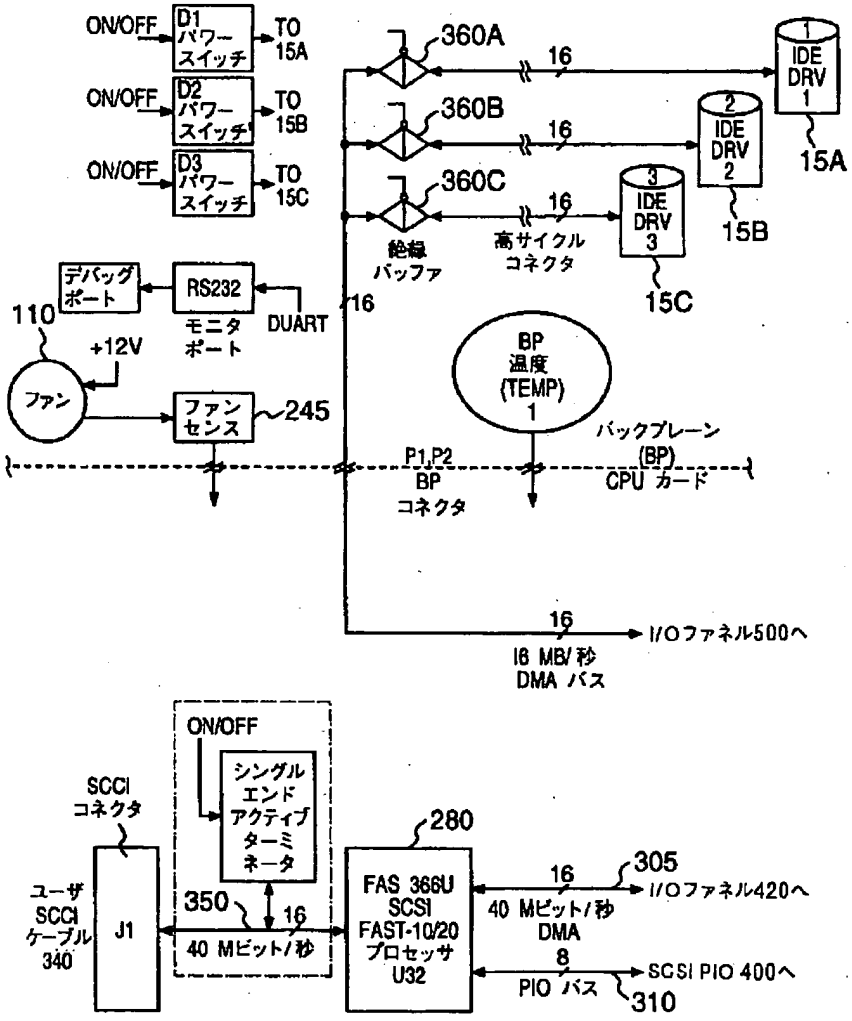
【図17】


クロック
600
SCSI 位相
605
DCM 位相
615
IDE 位相
610
時間

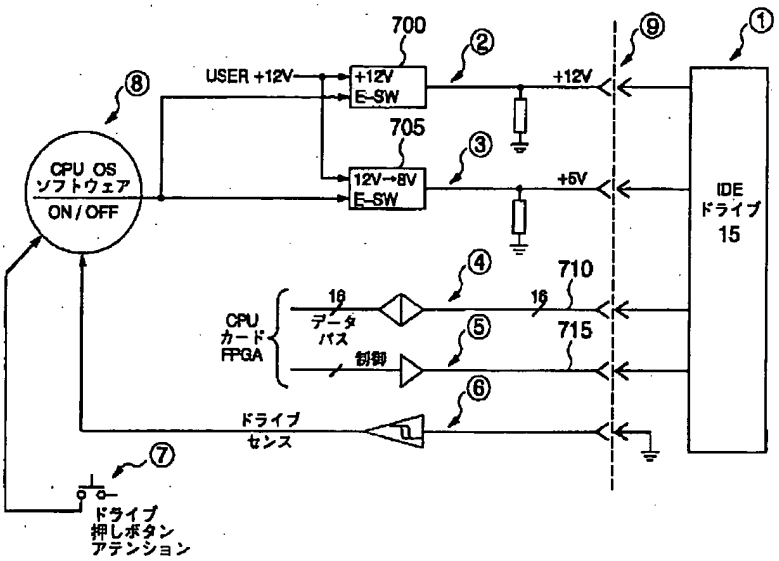
【図13】


1M x 32(4MB) 処理メモリ
275
270
40 MHz 32-BIT マイクロプロセッサ LR33310-40 U33
325
32データ
32アドレス
ISO バッファ
DRV 1 LEDS +PB
DRV 2 LEDS +PB
DRV 3 LEDS +PB
7777 ESD シールド
フラッシュブート +OS
T.O.D.
DUART
385
300
290
バッテリ
365
モニタポート
メンテナンスポート
SNS 0
SNS 1
CPU 周辺機器デコーダ
P(PEC)U80 レジスタ
397
399
MUX バッファ
RCA[9..0]
401
RCA
DPI
DPO
HWC RAC CAC
256 x 16 I/O コマンドキュー 40 MHz プロセッサ
415
ユーザ LED ステータス
システム電圧モニタ
電源投入シーケンサ
リセット
ディスクキャッシュメモリ
(DCM) 1M x 32 + P (4MB)
265
403
470
465
460
455
395
390
IDE ポインタ
SCSI ポインタ
DCM ポインタ
445
450
440
430
435
40 MHz 256 x 32 フレームベース I/Oキュー (160 MB/SEC)
475
480
485
425
FPGA ブート ROM U30
U1 FPGA (EPF10K30 BC356-4 BASED)
10 MHz
40 MHz 16↔32 I/O ファネル
500
端末センサ
ドライブ電力
420
SCSI PIO
400
260
360A～360C から
BP 温度 1
CPU 温度 0
ON/OFF
305 280へ
310 280へ

【図14】


ON/OFF
D1 パワースイッチ
TO 15A
ON/OFF
D2 パワースイッチ
TO 15B
ON/OFF
D3 パワースイッチ
TO 15C
360A
360B
360C
16
1 IDE DRV 1
15A
2 IDE DRV 2
15B
3 IDE DRV 3
15C
絶縁バッファ
高サイクルコネクタ
デバッグポート
RS232
モニタポート
DUART
110
+12V
ファン
ファンセンス
245
BP 温度 (TEMP) 1
バックプレーン (BP)
CPU カード
P1,P2 BP コネクタ
I/Oファネル500へ
16 MB/秒 DMA バス
ON/OFF
シングルエンドアクティブターミネータ
SCCI コネクタ
ユーザ SCCI ケーブル 340
J1
350
40 Mビット/秒
280
FAS 366U SCSI FAST-10/20 プロセッサ U32
305
I/Oファネル420へ
40 Mビット/秒 DMA
8
SCSI PIO 400へ
PIO バス
310

【図18】



フロントページの続き

(72)発明者　ホセ　プラトン　バスコ
アメリカ合衆国　フロリダ州　ウエリントン，ニアンティック　テラス　1276

(72)発明者　トマス　ウイリイ
アメリカ合衆国　フロリダ州　レイク　ワース，シダー　ハースト　コート　7627

【外国語明細書】

EXPRESS MAILING LABEL EM003714706US

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

APPLICATION

FOR

UNITED STATES PATENT

FOR

HIGH DENSITY RAID SUBSYSTEM WITH HIGHLY INTEGRATED CONTROLLER

Inventors:

Adriano Roganti
Thomas Wille
Ron Smith
Jose Basco

SPECIFICATION

FIELD OF THE INVENTION

The invention relates to disk drives, and more particularly to RAID array subsystems and controllers.

BACKGROUND OF THE INVENTION

Hard disk storage has become ubiquitous for virtually every personal computer and server, as well as many other related types of systems. In many instances, such storage represents the only repository for mission-critical information for at least the time between backups. As a result, these storage devices must be highly reliable and maintain extremely high data integrity.

Many types of storage subsystems have been developed to ensure against data corruption, including mirrored drives, failover systems, and multiply redundant drive subsystems. A form of multiply redundant subsystem which has become particularly well-regarded for its high reliability is the "redundant array of inexpensive

drives, " or RAID subsystem.

RAID subsystems typically have been implemented in servers and other computer systems. In general, RAID subsystems include two or more disk drives (typically of the same capacity, and frequently of the same type) and, in at least some forms of RAID implementations, are configured such that each drive serves as the primary storage device for a first portion of the data stored on the subsystem and serves as the backup storage device for a second portion of the data. Various backup schemes for RAID systems have been developed, including RAID 0, RAID 1, and RAID 5. In RAID 0, no data redundancy is provided, and the capacity of the RAID array is simply the sum of the capacities of the individual drives. In RAID 1, each drive is backed up by an associated drive much like mirrored drives. RAID 1 is implemented in most instances with even numbers of drives. RAID 5, on the other hand, can be implemented by a varying number of drives, typically beginning at a minimum of three (two drives would simply degrade to RAID1.) For a five-disk RAID 5 subsystem, each drive serves as primary storage for 80% of its capacity, and secondary storage for 20% of its capacity. As a result, the storage capacity of such an array is 80% of the sum of the capacities of the drives.

In general, prior art RAID subsystems have been external to the server. This has imposed space and reliability issues, among other things. Conventional sizes of PC cases typically offer only a very limited number of bays for disk storage, and conventional RAID arrays are simply too large to fit the available space. This imposes the requirement for extra floor space in what is typically already a crowded area, but also imposes the requirement for an external cable to connect the server or other PC to the RAID device. One of the more common causes of failures for external devices is cable failure, often due to human error in bumping or inadvertently disconnecting the cable.

In some instances, for example some models of the HP NetServer line, oversized cases have been provided which provide extra bays for storage devices. For example, the NetServer LM product includes a double-wide case with a RAID controller inserted into an expansion slot of the server and a stack of eight bays for drives conforming to the 3.5" form factor. However, this solution obviously requires buying a specific vendors specific model of server and thus limits the user's options. Moreover, the RAID controller occupies an expansion slot which might otherwise be available for other devices. These constraints of the prior art have left those wishing to include RAID subsystems in their existing servers with very limited options.

The assignee of the present invention has previously attempted to resolve some aspects of the dilemma presented to end-users attempting to include RAID

subsystems in their existing servers. For example, Aiwa/Core's MicroArray is a RAID subsystem configured to fit within the 5.25" full height form factor. This permits the subsystem to be installed within most existing cases and therefore avoids the footprint and external connection issues of other prior art. The MicroArray product permits a plurality of IDE disk drives (up to five) conforming to the 2.5" form factor to be inserted into the subsystem. The MicoArray product includes within its 5.25" form factor the RAID controller and related electronics necessary to interface the IDE drives to the RAID controller and to provide an external SCSI interface to the host system.

While the MicroArray product offered many advantages over existing prior art, it did have some drawbacks. One significant drawback was that it required the use of expensive 2.5" disk drives, which typically offer far less capacity and less reliability than drives conforming to the 3.5" form factor while at the same time costing significantly more. Because of these limitations, 2.5" drives have typically found a market only in laptop applications, while most desktop applications have used 3.5" drives.

In addition, the RAID controller of the MicroArray product offered limited throughput compared to that available in other devices today and comprised a complicated — and therefore expensive — design. The controller implemented substantially conventional wisdom and offered independent I/O channels for each of the drives in the array. This imposed significant space requirements which prevented the use of any drive larger than those complying with the 2.5" form factor.

As a result there has been a need for a RAID subsystem which is capable of fitting with a 5.25" full height bay of a conventional server case, while at the same time offering an integrated controller within that space and the use of low-cost, high capacity 3.5" drives.

SUMMARY OF THE INVENTION

The present invention describes a RAID subsystem which substantially improves upon the prior art in offering substantially improved capacities, improved throughput, higher reliability, and lower cost while still fitting within a single 5.25" full height bay. The RAID subsystem of the present invention includes the use of a plurality of 3.5" disk drives using the EIDE interface, while at the same time offering the Ultra-SCSI interface to the host system with its desirable high-speed data transfer rate.

To achieve the foregoing, careful management of the mechanical and

electrical interfaces has been required — both between the individual drives in the array and the controller, and between the subsystem and the host – to fit the desired capabilities within a tightly limited space. In additional, careful thermal management has been required because of the very limited availability of space for airflow within the subsystem. Finally, the foregoing requirements substantially prohibit the use of conventional controller designs, such that a highly integrated RAID controller has been developed as part of the present invention. The controller of the present invention has the additional feature of offering substantial benefits in areas outside the mechanical design of the present RAID subsystem.

In addition to the mechanical, electrical and thermal problems described above, the present invention is intended to permit ease of maintenance by the end user, which imposes the additional requirement of permitting the end user to have easy access to the drives integrated into the subsystem. This has been achieved by permitting the end-user to remove the front panel of the subsystem, which allows the end-user to remove one or more of the drives in the manner described in U.S. Patent Application S.N. 08/931,766, filed on 9/16/97 and entitled Disk Drive Latch, assigned to the assignee of the present invention and incorporated herein by reference. At the same time, the end-user's desire for information on the operation of each drive substantially demands that status and access information be delivered to at least the front panel of the subsystem. While the most reliable method for providing such information to the user is by integrating LEDs or other display devices into the printed circuit board on which the RAID controller is mounted, implementing such a design also imposes the limitation that the end user may also be able to touch at least an edge of that printed circuit board. This results in the requirement that the controller board be protected from significant amounts of electrostatic discharge, or ESD, in the event the end-user does not take adequate precautions while accessing the interior of the subsystem.

As noted previously, the controller of the present invention is subject to multiple design constraints not generally found within the prior art. Included in these are space limitations, in that the space available within the form factor for the controller board simply does not permit the use of conventional controller designs. Second, the thermal requirements imposed by the form factor reinforce that conventional controller designs are unacceptable as generating excessive heat. Third, cost requirements make the use of multiple controllers undesirable.

As a result, a highly integrated RAID controller has been developed in which a single I/O channel is provided for use by the SCSI host functions and the drives included within the array, as well as for DMA functions. The single I/O channel is

time-multipexed to permit each drive to access the controller for a prespecified, finite period, and also to permit the SCSI host portion of the interface to access the controller for a similar prespecified finite period. By the use of suitable clocking rates, the single-chip controller can thus attend to each of its required functions while at the same time managing the requisite DMA functions. In one embodiment, the engine of the controller may be implemented in an off-the-shelf field programmable gate array, or FPGA, although the design may also be implemented in an ASIC or other similar device. While the controller of the present invention is shown herein used with internal RAID subsystems, the design has application for both internal and external RAID subsystems and may also have application entirely outside the RAID environment.

In addition, the array of the present invention permits hot-swapping of disk drives maintained within the array. Activation of a drive-specific switch accessible to the user causes the firmware of the system to power down the drive. The drive may then be removed and a replacement drive installed. The firmware then automatically senses the installation of the new drive, and reapplies power as well as reconnecting data and control signals. The technique allows maintenance to be performed without down time or loss of data, suppresses power surges and provides protection from electrostatic discharges.

These and other features of the present invention will be better appreciated from the following Detailed Description of the Invention, taken in conjunction with the attached Figures.

BRIEF DESCRIPTION OF THE DRAWINGS

Figure 1 shows in front three-quarter perspective view the RAID subsystem of the present invention with the top cover removed.

Figure 2 shows in exploded front three-quarter perspective view the various components of the subsystem of the present invention.

Figure 3 shows in top plan view the RAID subsystem of the present invention with the top cover removed.

Figure 4 shows a front elevational view of the RAID subsystem of the present invention with the front cover removed.

Figure 5 shows a rear three-quarter perspective view of the subsystem of the present invention with the top cover removed.

Figure 6 shows a rear elevational view of the subsystem of the present invention with the rear cover plate removed, and in particular shows the backplane.

Figure 7 shows in front elevational view the layout of the backplane of the present invention.

Figure 8 shows in top three-quarter perspective view a single drive and associated mounting bracket with backplane interface board.

Figure 9 shows in perspective view the drive mounting bracket of Figure 8, and in particular shows the ribbon cable interface between the drive and the bracket.

Figure 10A shows in layout form one side of the RAID controller board.

Figure 10B shows in layout form the second side of the RAID controller board.

Figure 11 shows in schematic block diagram form the RAID controller of the present invention including the RAID engine.

Figures 13 to 16 show the internal configuration of the RAID engine of Figure 12.

Figure 17 shows the timing of various operations of the RAID engine shown in Figure 12.

Figure 18 shows in schematic form the hot-swap capability of the RAID array.

DETAILED DESCRIPTION OF THE INVENTION

Referring generally to Figures 1 through 8, and particularly to Figures 1 and 2, the RAID subsystem 10 of the present invention can be better appreciated. As will be appreciated better hereinafter, the top cover 12 of the subsystem has been removed in Figure 1 but is evident in Figure 2. A plurality of conventional IDE compliant disk drives 15A, 15B and 15C (where IDE includes within its general scope EIDE and Ultra DMA drives), each of which also complies with the accepted 3.5" form factor, are mounted behind a front bezel 20 and within a case 25. The case 25 cooperates with the bezel 20 to fit within the conventional 5.25" full height form factor, which is generally accepted as approximately 5.25" wide by 3.25" high. A latch 22A, formed integrally with the bezel 20 and mated to a receiver 22B in the case 25, cooperates with L-shaped posts (not shown) on the inside of the opposite end of the bezel which engage the inside of the case 25 to permit the bezel to be unlatched, swung out and removed for maintenance. The length of the form factor is less tightly controlled but is generally on the order of eight to ten inches. An internal top plate 30 and internal side wall 35 are rigidly affixed to the case 25 to define a first cavity 40 suitable for mounting the 3.5" drives 15A-C. The top plate 30 and side wall 35 also enclose a second, long, narrow cavity 45 to the left of the first cavity 40, the use of which is discussed in greater detail hereinbelow.

Each drive 15A-C is mounted within a U-shaped drive bracket 50 (best seen in Figures 8 and 9 and described in detail in connection therewith) which comprises

a pair of rails 55A-B and a drive extension board 60. A mounting mechanism 65 is mounted on the rails 55A-B, which mechanism is better described in U.S. Patent Application S.N. 08/931,766, filed on 9/16/97 and entitled Disk Drive Latch, referred to hereinabove and incorporated herein by reference. The rails 55A-B slidably fit within grooves 70 in matching mounting plates 75A-B (best seen in Figures 2 and 4), which are affixed to the interior of the right sidewall 80 of the case 25 and the right face of the internal sidewall 35.

Positioned behind the drive extension boards 60 associated with each of the drives 15A-C is a backplane 85, described hereinafter in connection with Figures 6 and 7. The backplane 85 includes a plurality of connectors 90 (shown particularly in Figure 2 and Figure 7) to mate with a matching connector 90A on each of the drive extension boards, and also includes a connector 95 (best seen in Figures 6 & 7) for mounting to a RAID controller printed circuit board 100 mounted within the cavity 45 down the left side of the case 25. A rear cover plate 105 is affixed to the rear of the case 25 to enclose the backplane 85 and the back edge of the RAID controller board 100, and supports a fan 110. The rear cover plate 105 is spaced behind the backplane 85 to form a plenum chamber 115 to permit the fan to cool efficiently the RAID controller board 100 and the drives 15A-C in the tight spacing imposed by the case 25. Other details of the various elements mentioned above will be described in connection with other Figures.

Still referring generally to Figures 1-8 and with reference particularly to Figure 3, the arrangement of the disk drives 15A-C and their connection to the backplane 85 can be better appreciated. The drives 15A-C (only drive 15A is shown in Figure 3) are latched into the case 25 by virtue of latching mechanism 65, which urges the connector 90A affixed to the drive extension board 60 into mating contact with the connector 90 on the backplane 85. It will be appreciated that the drive extension board 60 is spaced somewhat behind the drive 15A to permit, among other things, variations in the length of the drives 15A-C and also to provide an airflow chamber. Likewise, the spacing of the connectors 90 and 90A creates an airflow chamber 150 between the drive extension board 60 and the backplane 85. The drive 15A can be seen to be connected to the drive extension board by a flexible ribbon cable 60A, visible here but better seen in Figure 9. The ribbon cable 60A connects to the IDE connector included with the drive 15A, and allows for slight variations in the location of the connector on different types of drives.

The backplane 85 is affixed to the case 25 by virtue of an upper and lower pair of mounting brackets 155 (at the left) and another pair 160. The mounting brackets 155, which are, in an exemplary embodiment, integrally formed with the internal side

wall 35, may be seen to be double bent. Mounting brackets 160 may be seen to be affixed to the sidewall 80. While not necessary in many cases, the additional resiliency offered by the double bend in mounting bracket 155 aids in absorbing the deflection forces imposed on the drive and the backplane by the insertion and removal process. In addition, the resiliency of the mounting brackets and the backplane, as well as the ribbon cable 60A, are believed helpful in isolating the drives from any vibration imposed by the fan, the remaining drives or elsewhere in the system. The combination is believed helpful in increasing the reliability of the system and extending the life of the drives. In at least some instances, the flexibility of the backplane 85 and the drive extension board 60, together with the ribbon cable 60A, will be sufficient to provide adequate resiliency and isolation.

The plenum chamber 115 may also be appreciated from Figure 3, and can be seen to form a decompression space in front of the fan 110. The plenum chamber 115 collects air drawn around the drives 15A-C through cavity 40 and collected in cavity 150 as well as air drawn past the RAID controller board 100 through cavity 45. The spacing between the backplane 85 and rear cover plate 105 can be adjusted as necessary to optimize the efficiency of the fan 110 in drawing air through the RAID array and maintaining the array within an acceptable thermal range.

For ease of manufacturing, the RAID controller board is slidably mounted within the cavity 45. Two pairs of guides 165, which may be formed unitarily with the top wall 30 by being punched downward essentially to form a slot, position the top edge of the board 100 centrally within the cavity 45, in combination with a similar slot (not shown) formed in the bottom of the case 25. A similar guide 170 may also be provided at the front of the wall 30.

Referring next to Figure 4, the stacking arrangement of the drives 15A-C can be better appreciated as well as the airflow through the cavities 40 and 45. As with Figure 3, the top cover is not shown. The RAID controller board 100 can be seen to be centrally located in the cavity 45, permitting airflow past either side of the board 100. In addition, the gaps between the mounting blocks 75A-B and the rails 70 can be seen to provide air passages past either side of the drives 15A-C within the cavity 40. By properly sizing the fan 110 and plenum chamber 115 to match the airflow through the cavities 40 and 45, sufficient cooling is provided to the drives and to the RAID controller board to permit long-term continuous operation. It will be appreciated that additional drives may be included in the event thinner drives are used, with commensurate changes to the RAID controller discussed in connection with Figure 11.

In addition, a leaf spring 175 may be positioned at the front of the cavity 45

both to urge the board 100 into the proper position and also to provide a ground plane connection to the board 100 for discharging any electrostatic charge which might be imposed on the board by the user during maintenance of the array. It will be appreciated that, unlike most subsystems within the computer system, the front edge of the RAID controller 100 will be accessible to the user from the front panel of the computer system simply by removing the bezel 20. As a result, a suitable path to ground for ESD purposes is appropriate, and can be provided by plating with a conductive material a portion of at least one side of the board 100 near its front edge and connecting that plating to the case through the leaf spring 165. The leaf spring is typically constructed of copper or other suitable spring material. Copper plating of such other materials may be desirable in at least some instances. The plating of the board 100 may best be seen in Figure 10A, where the plating is identified by reference numeral 285.

Further shown in Figure 4 are a pair of LEDs 180A-B for each drive, together with a pushbutton 185 for each drive. The LEDs 180A typically indicate status of the associated drive and may be multicolor LEDs which use different colors to indicate different operational states. The LEDs 180B typically indicate activity on the associated drive. The pushbuttons
185 are used to signal the RAID controller that the user desires to disconnect the associated drive. By depressing the pushbutton 185, the RAID controller disconnects power and signal paths to the associated drive, allowing that drive to be safely removed while the remainder of the array continues to operate. Once the drive has been electrically disconnected from the array, the drive may be physically removed by virtue of latches 65. That drive or another equivalent drive may then be added back into the array by fastening latches 65. In an exemplary embodiment, insertion of the drive connector 90A into the backplane connector 90 causes the addition of the drive to be sensed by the array; however, in some embodiments the array may be caused to sense the addition of the new drive by again pushing the associated pushbutton 185.

With reference next to Figures 5, the rear portion of the subsystem can be seen in a top three-quarter perspective view, such that the cooling fan 110 and external connectors for connecting the subsystem to the host can be better appreciated. As discussed in connection with Figure 3, the cooling fan 110 is positioned centrally behind the backplane and spaced therefrom sufficiently to avoid unacceptably turbulent airflow through fan, which increases the amount of airflow past the drives and printed circuit boards and therefore optimizes the cooling effects of the fan. At the left of the fan 110 is positioned a nine-pin D-shell connector 200,

typically used to connect to a monitoring device such as the ArrayView product offered by the assignee of the present invention or other suitable device for monitoring the status of the subsystem. Below the D-shell connector 200 is a conventional power connector 205. The D-shell connector 200 and the power connector 205 are, in the exemplary embodiment described herein, connected to the backplane 85 and extend through openings in the rear cover plate 105. At the right side of the fan is a high density connector 210 conforming to the single-ended Ultra-Wide SCSI standard, together with a suitable jumper block 215 for setting the ID of the unit, performing various diagnostics, and other conventional functions. The connectors 210 and 215 are typically affixed to the RAID controller board 100, and extend through openings in the rear cover plate 105. The SCSI connector 210 typically provides the interface to the host system, and the entire subsystem appears as a single SCSI device to the host adapter in the host system. In other embodiments, the subsystem may comply with different interface standards such that different connectors may be offered, including differential SCSI, wide SCSI, or some other interface.

Reference is next made to Figure 6, which shows the array subsystem in rear elevational view with the rear cover plate (including the fan) removed and thus shows in detail the layout of the back of the backplane 85, and Figure 7, which shows in front elevational view the layout of the backplane 85. With particular reference to Figure 6, the connectors 200 and 205 can be seen to be integral with the backplane 85, and the manner by which the dual connector 95 connects the backplane 85 to the RAID controller board 100 can also be seen. In addition, a variety of vents or cutouts 225, both at the periphery and through the backplane 85, can be seen to exist in the backplane to improve airflow into the plenum chamber 115. The backplane is held in place by four screws (not shown) which pass through holes 230 and mount into the mating pairs of mounting brackets 155 and 160. Also mounted on the backplane 85 is an alarm 235 which responds to signals from a variety of sensors which monitor array performance, including for example one or more drive temperature sensors 240, a fan sensor 245, and so on, which in the exemplary embodiment shown herein may be mounted on the front of the backplane as shown in Figure 7. In addition, the connectors 90 shown on the front of the backplane are typically high cycle, low insertion force connectors which provide both a conventional IDE bus and power to the associated drive. The drive extension 60 then provides the appropriate mechanical interface to the drives, including conventional IDE connectors and conventional power connectors. Although the particular ordering of the drives 15A-C which plug into the connectors 90 is not critical, in the exemplary embodiment

described herein the drive associated with the top connector is typically assigned drive 0, the middle connector drive 1, and the bottom connector drive 2.

Referring next to Figures 8 and 9, the manner by which a single drive 15A fits into a drive bracket 50 can be better understood. The drive bracket 50 comprises, as noted above in connection with Figures 1 and 2, a pair of rails 55A-B together with a drive extension board 60. The drive 15 is mechanically affixed to the bracket 50 by means of conventional machine screws, and electrically connects to the bracket through the cable 60A and the connector 90A to the backplane connector 90, as well as through a conventional Amphenol power connector 60B. The latch mechanisms 65 may also be appreciated.

Referring next to Figures 10A and 11, the layout of the RAID controller board 100 may be seen. The RAID controller board 100 comprises a single double-sided printed circuit board, the schematic of which can be better appreciated from Figure 12, discussed below. Viewed from the side shown in Figure 10A, which can be seen to be the outboard side, the connectors 210 and 215 may be seen at the far left. The RAID controller includes a RAID engine integrated circuit 260 (which may be either a Field-Programmable Gate Array, an ASIC or other suitable implementation), to perform the necessary queuing and DMA functions. The RAID engine 260 communicates with cache memory 265 (Figure 11), a RISC CPU 270 for managing the operation of the RAID controller, its associated CPU memory 275 (both Figure 11), and a SCSI processor 280 (Fig. 10A) for managing the host interface. The LEDs 180A-B and pushbuttons 185 can be seen to be connected to the RAID controller board at the forward edge (Figure 11) while on the opposite side of the board the conductive ESD plating 285 (discussed generally in connection with Figure 3) may be seen. The exemplary embodiment of the RAID controller board 100 shown herein also includes a pair of connectors 275 for permitting the backplane 85 to be connected into the board 100. A time of day/date chip 290 may also be provided as well as various other sensors and logic which perform conventional functions as described in connection with Figure 12. From the arrangement of the plating 285, it will be particularly appreciated from Figure 10B that a user performing maintenance on the subsystem of the present invention is substantially prevented from damaging the RAID controller as the result of any electrostatic charge the user may carry when performing otherwise acceptable maintenance, because the plating 285 is connected directly to the ground plane as discussed above.

Referring next to Figures 12 and 13 to 16, the electrical operation of the invention may be better understood. In general, the RAID subsystem appears to the host system as a single volume which externally complies with conventional SCSI

commands, but internally operates as a full RAID array. The RAID array operation is controlled by the RAID controller, which in turn operates by using time-division multiplexing and separate 32-bit DMA and CPU software process memory to allow for simultaneous non-contending activities at the engine's peak rate. The DMA or cache memory 265, which may for example be four megabytes configured as 1x36 memory, provides a single-cycle paged EDO pipeline with bandwidth on the order of 160 MB/sec. The CPU memory 275, which may be configured in an exemplary embodiment as four megabytes configured as 1x32, provides a two-cycle paged EDO pipeline with 80 MB/sec bandwidth.

The CPU 270, which may for example be an LSI LR33310-40 32-bit RISC processor operating at 40 MHz, cooperates with a FLASH ROM 300 which stores an embedded RAID operating system. At the center of the architecture is the RAID integrated circuit 260, which may for example be an Altera Field Programmable Gate Array (FPGA) or equivalent or may be configured as an ASIC, which provides command queues for each DMA I/O channel, manages the various data I/O queues, manages the bus activities on the key buses associated with it, and supports system peripheral functions. Five major buses are associated with the FPGA 260: a 40 MB/sec, 16-bit SCSI processor bus 305 (typically configured for Ultra-SCSI operation although other SCSI protocols can be supported); a 3.33 MB/sec 8-bit SCSI chip pipelined I/O bus 310; a 16 MB/sec 16 bit IDE drive bus 315; a 160 MB/sec 36-bit disk cache memory (DCM) bus 320; and an 80 MB/sec 32-bit CPU bus 325. The FPGA 260 is configured to permit operation of all five buses in parallel, with the RAID engine 260 operating at a sufficient speed to multiplex the access to the RAID engine 260 by the SCSI processor bus 305, the IDE bus 315, and the DCM bus 320 by allocating, within a defined cycle, one time slot for each of the SCSI processor bus 305 and the IDE bus 315, and two time slots for the DCM bus 320. In the exemplary embodiment discussed herein, a complete cycle may be on the order of 100 ns, with each of the four time slots allotted 25 ns. Because of the performance of the RAID engine 260, the net subsystem throughput is primarily dependent on four factors: the performance of the IDE drives, the RAID function overhead in the embedded operating system, the performance of the user's host adapter, and the driver overhead of the user's host application.

Still referring to Figure 12, the operation of the system is substantially as follows: On power-up, the system comes to a stable state by loading the operating system from the Flash ROM 300 into CPU memory 275 associated with the RISC processor 270 and initializing the remainder of the system to known states. At some point after initialization, a request either to read or to write will be received from the

host system 340 at the host SCSI bus 350, which may be terminated by a termination block 355 if appropriate. The request is then handled by the SCSI processor 280, which sends the appropriate signals to the RAID Engine 260 over a pipelined I/O bus 310 and receives back the appropriate confirmation signal. The data is then made available by the host system over the SCSI DMA bus 305. At this point the Disk Cache Memory 265 is empty. If the request is to write information, the CPU 270 instructs the RAID engine 260 to pass the data to the DCM 265, where it can be maintained in cache. Thereafter, during background processing, the data can be written to assigned disk(s) as appropriate (in accordance with the RAID striping being used) by first having the data accessed by the RAID engine 260 over the bus 320 and written out over the IDE bus 315 to ISO disk buffers 360A-C. The data is then written to the specific disks 15A-C. It will be appreciated that the bus 320 comprises, in the exemplary system described herein, ten address lines and 36 data lines. Likewise, the bus 325 comprises thirty-two address lines and thirty-two data lines. The process ends with a confirmation signal supplied from the RAID engine to the SCSI processor 280 and thence to the host 340. The timing of the various events will be discussed in greater detail in connection with Figures ~~12~~ 13 to 16 and ~~13~~ 17. In figure 13, DPI and DPO represent "data path in" and "data path out", respectively.

In a read operation, the process is substantially similar, though somewhat reversed. The process begins by enabling the SCSI interface to be active, typically done at startup. The host then sends a confirmation/acknowledge signal and executes a set-up, followed by sending a request for specified data over the PIO bus 310. The request is then detected by RAID engine 260, which passes it to the drives 15. The data is returned from the drives to the RAID engine 260, where it is passed to the disk cache memory 265 for interim storage. At the appropriate time, the CPU 270 causes the data to be read from the DCM 265 via the bus 320 and passed through the engine 260 to the SCSI processor 280 over the data bus 305. The data is then passed from the SCSI processor 280 to the host over the bus 350.

In addition to its data handling functions, the RAID engine also manages a number of peripheral housekeeping functions. Included among these are monitoring of the over-temperature detector 240 and the fan error detector 245, generating alarm signals (when appropriate) at the alarm 235. The time-of-day/date clock 290 is also monitored, for which power is supplied by a battery or other power source 365 when the system is off. Hardware sense lines 370 can be monitored by means of status registers 375. Power surge control for the drives can be monitored at buffer 380. Monitoring of the subsystem is also provided over a duart 385. Typical monitoring is performed over RS232 links 390 for both monitoring and maintenance.

Referring next to Figures ~~12~~ 13 to 16 and ~~13~~ 17, the details of the operation of the RAID

engine 260 can be better appreciated, including the timing by which the signals on the various buses are multiplexed by the RAID engine. As with the prior figures, like elements have been assigned like reference numerals from Figure 12.

As before, when the host system is turned on, the RAID subsystem 10 initializes and the RISC CPU 270 generates a series of enabling factors as established by the software maintained in the FLASH ROM 300. The enabling factors place the IDE drives in known states and also place the SCSI processor 280 in an active and enabled state, including notifying the host system 340. The host system confirms and acknowledges the notification from the SCSI processor. In addition, the enabling factors place the RAID engine 260 in a known state, and in particular initialize a 40 MHz I/O Command Queue Processor 390 which is internal to the RAID engine 260. Figure 15 shows an alternative arrangement of a portion within a dot-and-dashed line in Figure 14.

After initialization, the host system sends data to be written to the drives 15A-C, as before. The information, which comprises header information and data, is supplied to the SCSI processor 280 over the bus 350. After processing by the SCSI processor 280, the header information is supplied to the RAID engine 260, indicated by the dashed line in Figure 13, over the eight-bit programmable I/O bus 310. The header information is supplied to the RAID engine 260 through a SCSI PIO 400, which has a data path in 405 and a data path out 410. The data path in 400 links to a one side of a mux 415, which in turn feeds, indirectly, the input to the I/O command queue processor 390. The I/O command queue processor 390 is a frame-based script processor and supplies half-word commands, row address commands and column address commands to a register 395 via a 16-bit bus. The register 395 can also receive addresses from the RISC processor 270; the processor 270 can also supply addresses to the DCM 265 via a 10-bit branch 397 of the bus 325, fed through a mux buffer 399. The output of the register 395 can be supplied via a multiplexed 10-bit bus (addressing one megabyte of address space) to the DCM 265 through a pipeline register 401 and buffer 403. The output of the register 395 provides, indirectly, the data-path-out referred to elsewhere in connection with the RAID engine 260, including the second input to the mux 415.

Concurrently with the header information supplied from the SCSI processor 280 to the SCSI PIO 400 on the bus 310, the data from the SCSI processor 280 is supplied to the RAID engine 260 via a sixteen-bit DMA bus 305. In particular, the data is fed to a 16-bit-to-32-bit funnel 420, operating at 40 MHz, because the RAID engine 260 operates internally at 32-bit width. The data is supplied to one side of a funnel mux 425 and then to one side of a I/O queue mux 430. The output of the queue mux 430 is supplied to a frame-based I/O queue 435, operating at 40 MHz

and configured at 256 x 32 to provide 160 MB/sec throughput. Other inputs to the I/O queue 435 include various IDE pointers 440, SCSI pointers 445 and DCM pointers 450. The data is clocked through to the output of the I/O queue 435 and supplied to a first pipelined output register 455, and then to one side of a DCM mux 460. The output of the mux 460 is provided to a second pipelined output register 465, through a buffer 470 and then out of the RAID engine 260 to the DCM 265.

The data is stored in the DCM 265 until appropriate for writing to appropriate ones of the disks 15A-C, typically determined by the RAID operating system according to conventional algorithms. At that time, the I/O command queue 390 issues a command to write the data to the disk drives. The data is supplied by the DCM 265 to a buffer 475 and then to a pipelined input register 480. The data is then provided to a second input register 485 as well as one side of a processor input mux 490. To write to the drives, the data is fed through the register 485 to the other side of the mux 430, and then to the I/O queue 435.

The data out of the I/O queue 435 is provided to the SCSI I/O funnel 420, but is also provided to a disk I/O funnel 500. The disk I/O funnel 500 reconverts outgoing data from a 32-bit data width to a 16-bit data width for communication with the disk drives 15. The remainder of the communication to the disk drives is as described in connection with Figure 12.

Retrieving data from the RAID subsystem is the other operation typically required of the RAID subsystem 10 by the host system 340. Retrieving data is initiated from the host system 340, which again supplies the host's request to the RAID engine 260 via the PIO bus 310, the SCSI PIO 400, and then the data-path-in 405 to the I/O command processor 390 through the mux 415. The I/O command processor 390 then supplies the appropriate RAC/CAC addresses via the register 395 to cause the data to be retrieved.

The appropriate addresses for the data desired by the host system are supplied to the DCM 265. If the data is maintained in the DCM 265, it is supplied via the registers 480 and 485 to the mux 430 and then to the I/O queue 435. From the I/O queue 435 the data is supplied to the SCSI I/O funnel 420 where the outgoing data is converted to 16-bit width. The data is then supplied on the DMA bus 305 to the SCSI processor 280, and finally out to the host 340 over the bus 350.

However, if the data requested by the host is not currently maintained in the cache 265, the data must be requested from the disks. In this instance, the addresses for the requested data are supplied via the registers 480, 485 and mux 430 to the I/O queue 435. The output of the I/O queue 435 is then supplied to the disk I/O funnel 500 and out to the drives 15A-C. The data is then retrieved from the

drives after the required latency, after which the data incoming from the drive is converted from 16-bit width to 32-bit width in the disk funnel 500. The output of the funnel 500 data is then supplied to the second side of the mux 425, and from there to the I/O queue 435 through the second mux 430.

The output of the I/O queue 435 is then fed through the I/O funnel 420 in the same manner as described above for outgoing data, with the result that the data is supplied to the host system in the conventional manner. For implementations based on a field programmable gate array (FPGA), an FPGA boot ROM 492 may be provided to personalize the FPGA upon power-up. In ASIC or other gate array implementations, such a boot ROM is not necessary. Likewise, the RAID OS is loaded into CPU memory 275 upon power-up, and all software control is derived from the instructions stored in the CPU memory.

A key feature of the operation of the present invention is that a single RAID engine 260 is able to manage the multiple DMA queues necessary to communicate with the SCSI processor, the disks, and the processor 270. This objective is achieved by time multiplexing the key buses which provide data to the RAID engine 260. This is possible because the I/O queue operates at an effective rate of 160 MB/sec, compared to the other devices which operate at no more than 40 MB/sec. This allows the RAID engine to allocate approximately one fourth of its time to each of the SCSI processor and disk drives, and to allocate approximately one-half its time to DMA addressing. Shown as Figure 17 is a timing diagram which provides the phased accessing necessary to provide the time division multiplexing important to some aspects of a presently preferred embodiment of the invention.

In particular, the 40 MHz clock of the I/O queue 435 and script processor 390 is shown at 600, while the SCSI Phase for access to the RAID engine is shown at 605. The IDE phase is shown at 610, while the DCM phase is shown at 615. In an additional feature, in the event that a cycle occurs when no access is requested from the IDE drives, the phase is reassigned for use by the DCM. Similarly, for those cycles in which the SCSI processor requires no I/O access the phase allocated for the SCSI processor is reassigned to the DCM. It will thus be appreciated that extremely high throughputs can be achieved with the present design.

Referrring next to Figure 18, the hot-swapping arrangement of the present invention -- by which one or more drives may be removed while the remainder of the array continues to operate -- may be better understood. In particular, in the event the user desires to remove one of the drives 15A-C, for example due to the failure of a drive, the user actuates pushbutton switch 185 associated with the drive to be removed. This signals the CPU 270, which operates under software control to signal

the FPGA control logic to power down both the 12 volt and 5 volt supplies 700 and 705, respectively, of the associated drive. In addition, the data path 710 and control path 715 are caused by the processor to be electrically disconnected from the remainder of the subsystem. At this point, the user can readily undo the latches 65 and remove the necessary drive.

To reinstall the drive, the user simply reverses the mechanical portion of the process by inserting the drive into the drive bay and latching the latches 65. A bistable latch senses the reinsertion of the drive, and signals the CPU 270 to reapply both power and signal connections to the newly-installed drive. In this manner the old drive may be removed and the new drive installed.

It can therefore be appreciated that a new and novel system for a RAID array subsystem and highly integrated controller has been described. It will be appreciated by those skilled in the art that, given the teachings herein, numerous alternatives and equivalents will be seen to exist which incorporate the invention disclosed hereby. As a result, the invention is not to be limited by the foregoing exemplary embodiments, but only by the following claims.

****

WE CLAIM:

1. A mass storage array subsystem comprising

a first longitudinal cavity for housing a mass storage controller which includes a SCSI interface to a host sytem and an IDE interface to internal mass storage devices,

a second cavity for housing a plurality of IDE drives capable of communicating with the host system such that the longitudinal axis of the first cavity is parallel to the longitudinal axis of the second cavity,

a backplane having vents therethrough and adapted to connect to each of the IDE drives housed within the second cavity, and

a plenum chamber formed by the backplane and the case to provide efficient airflow through the first and second cavities.

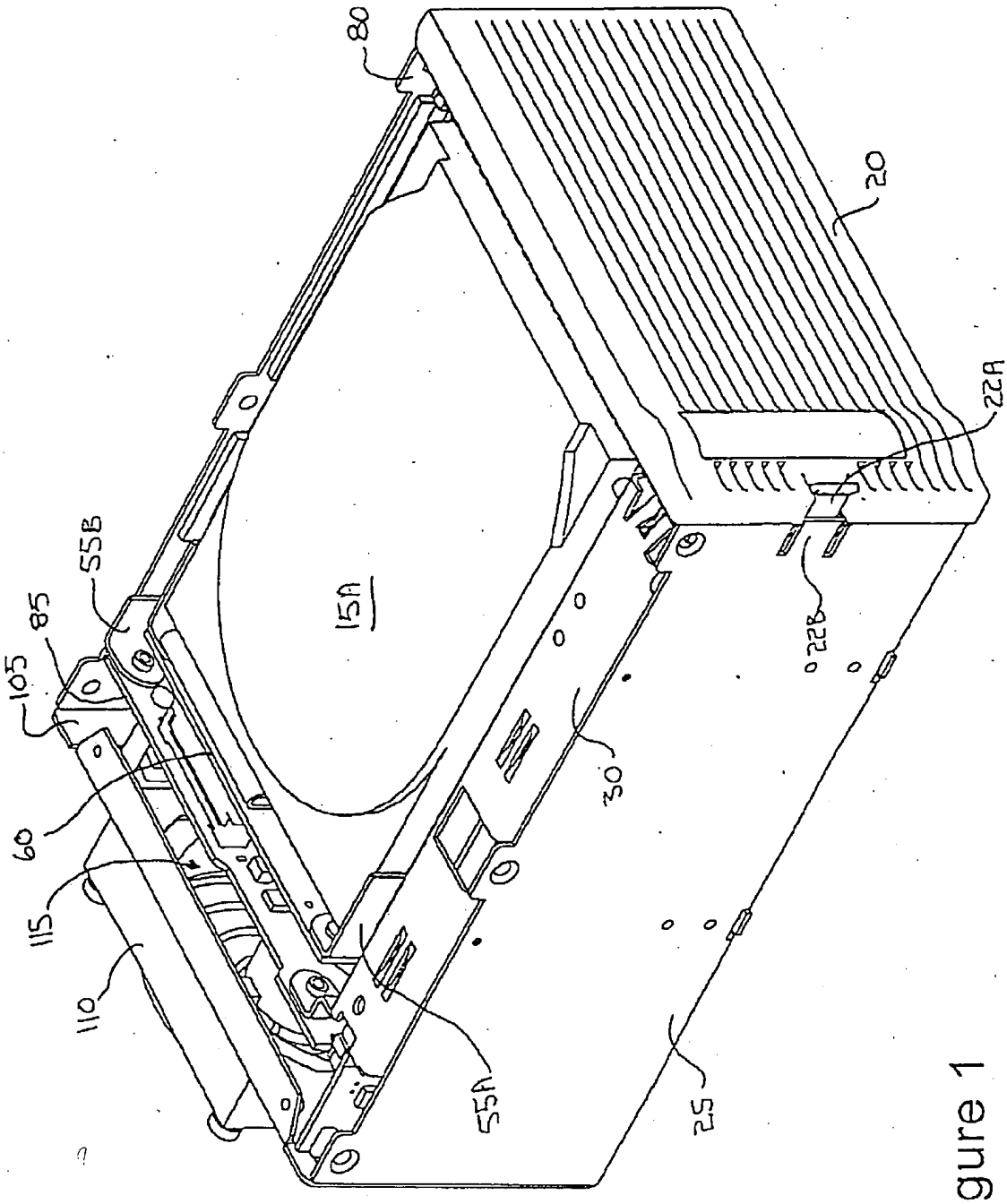


Figure 1

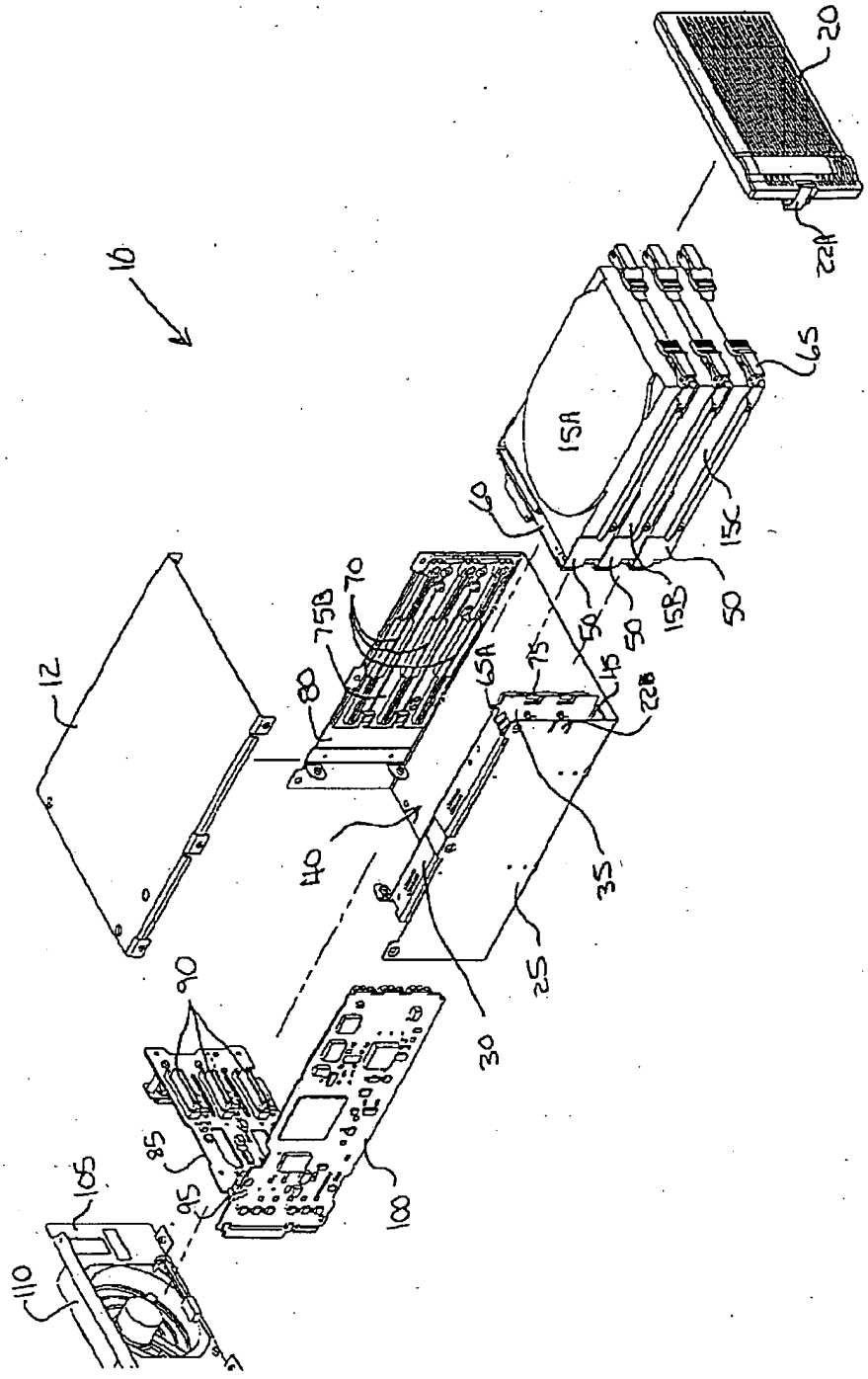


Figure 2

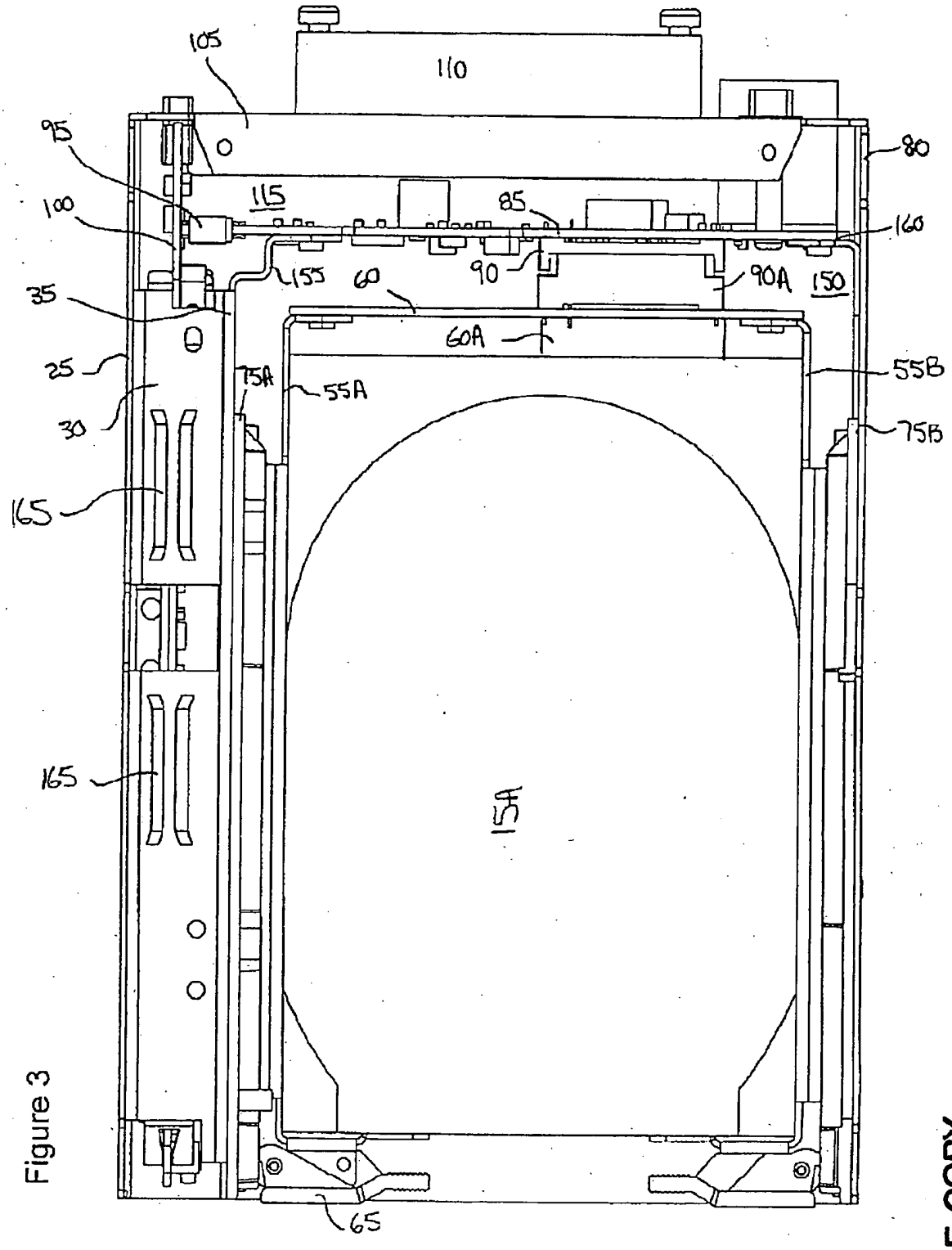


Figure 3

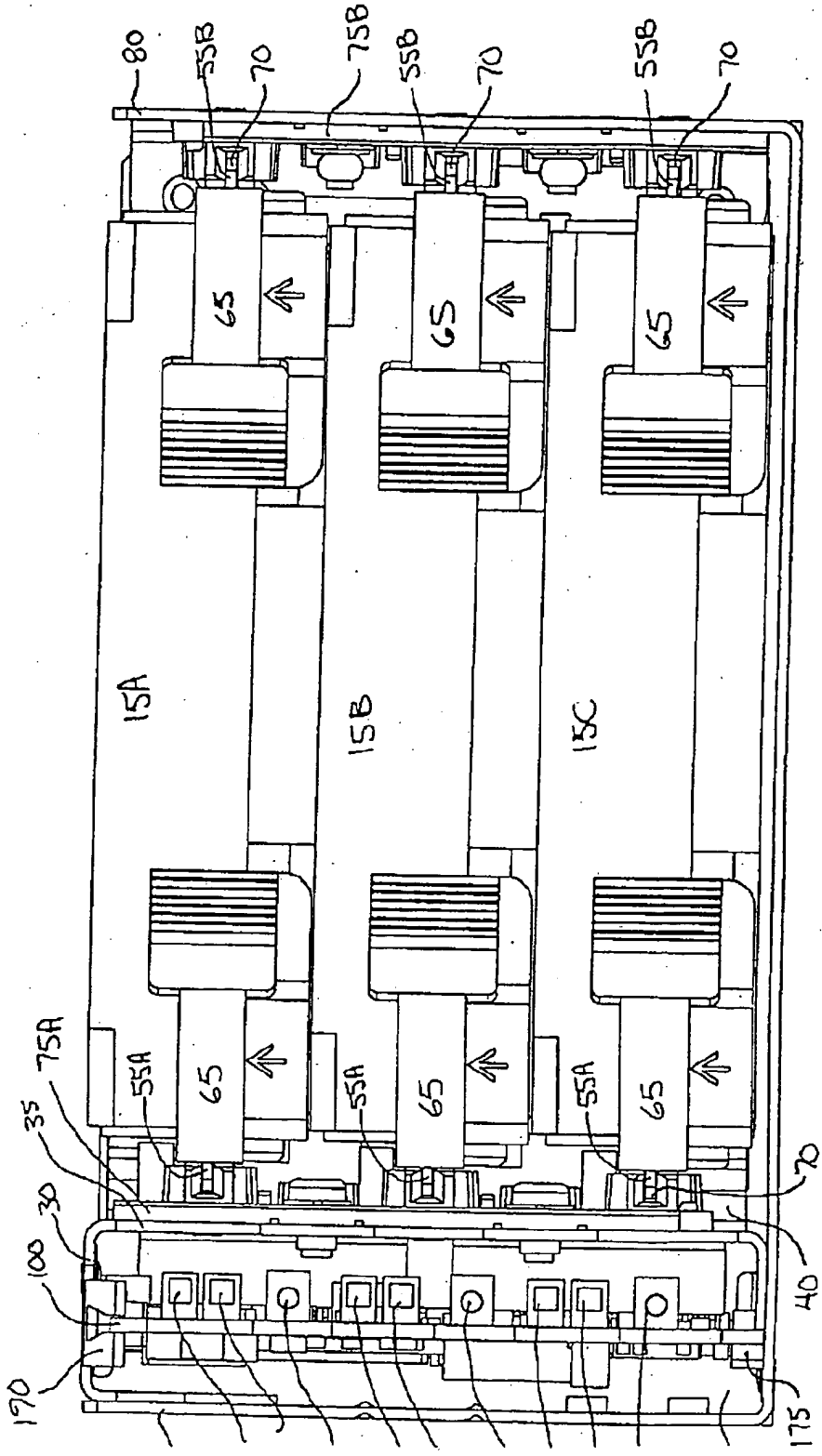
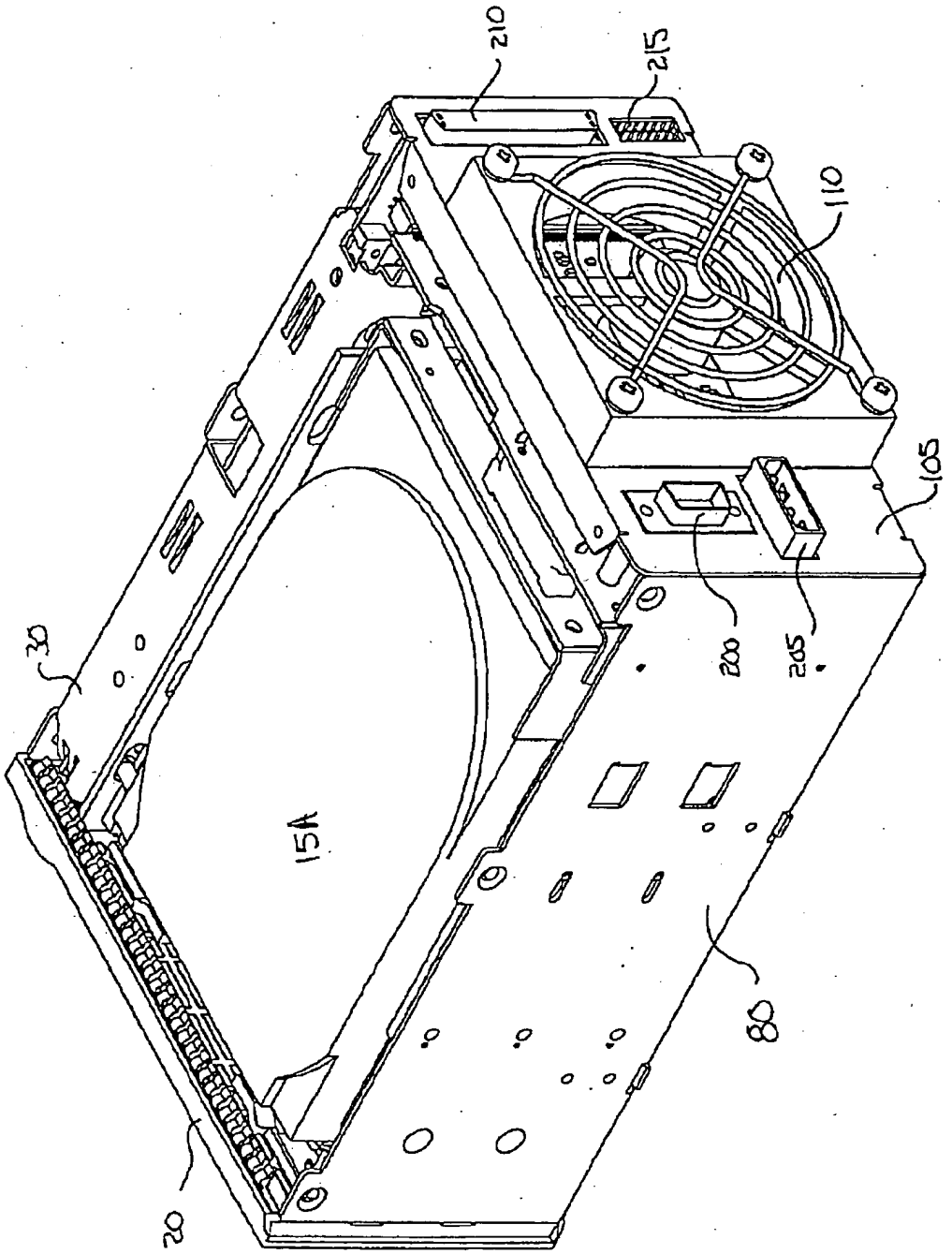


Figure 4

Figure 5



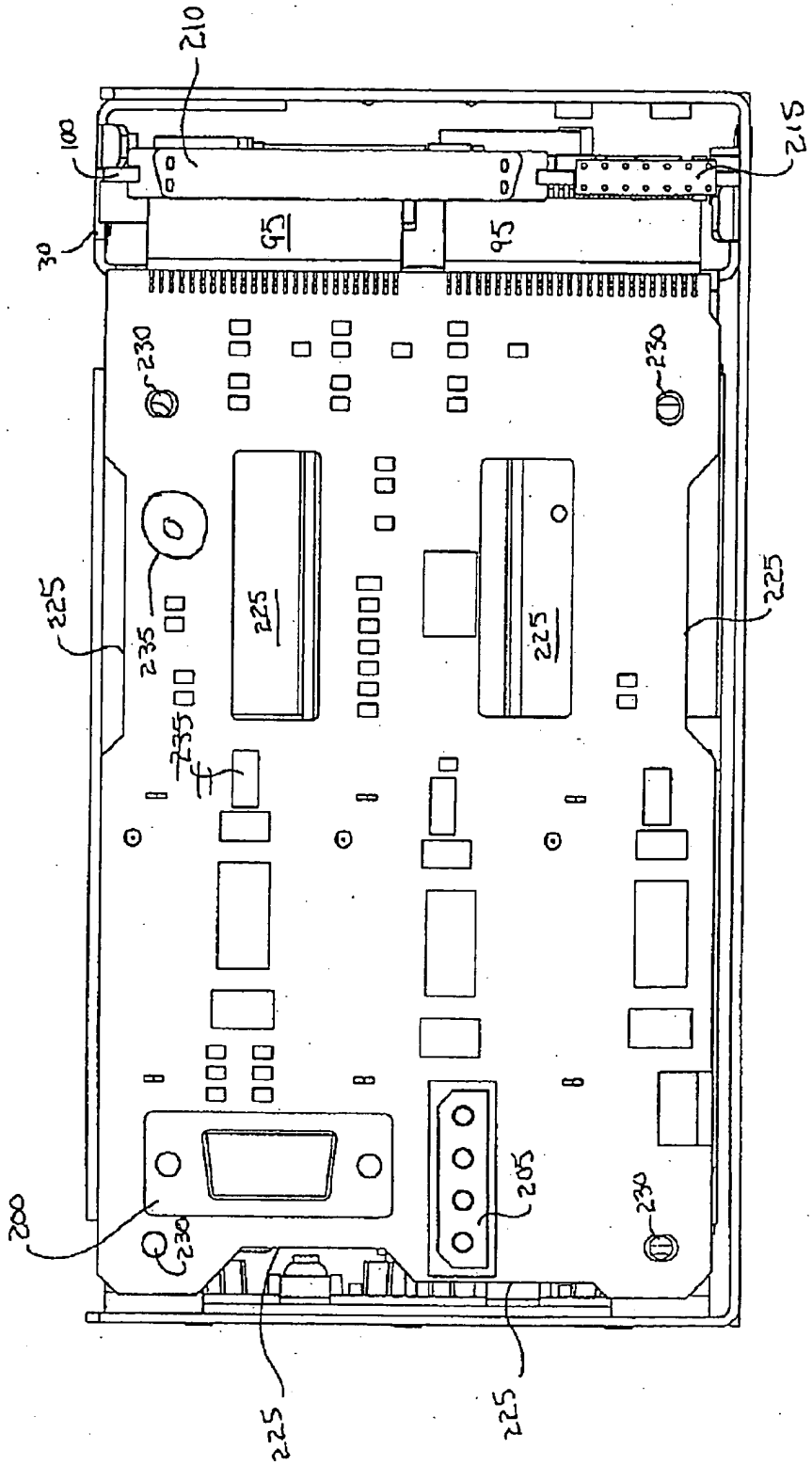


Figure 6

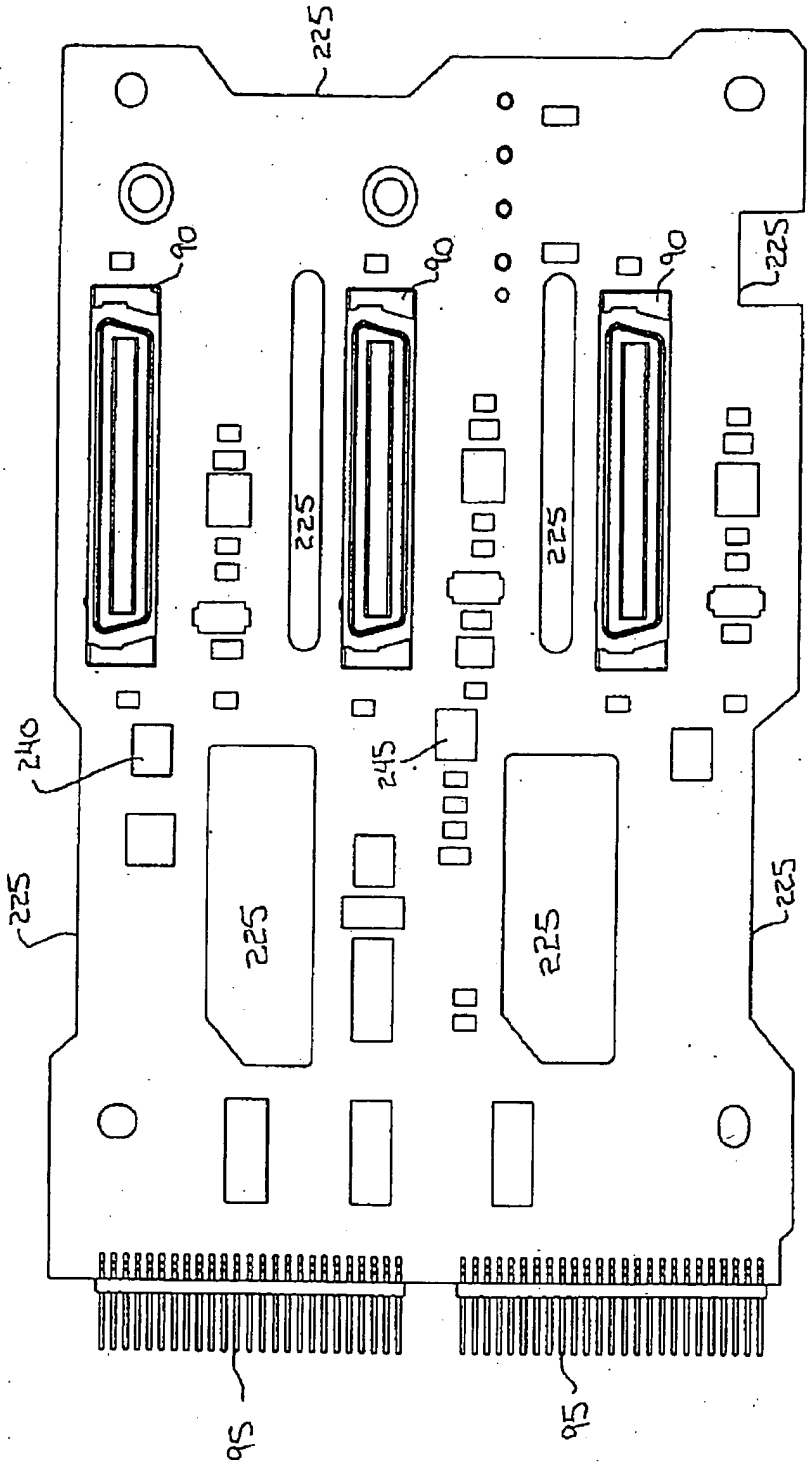


Figure 7

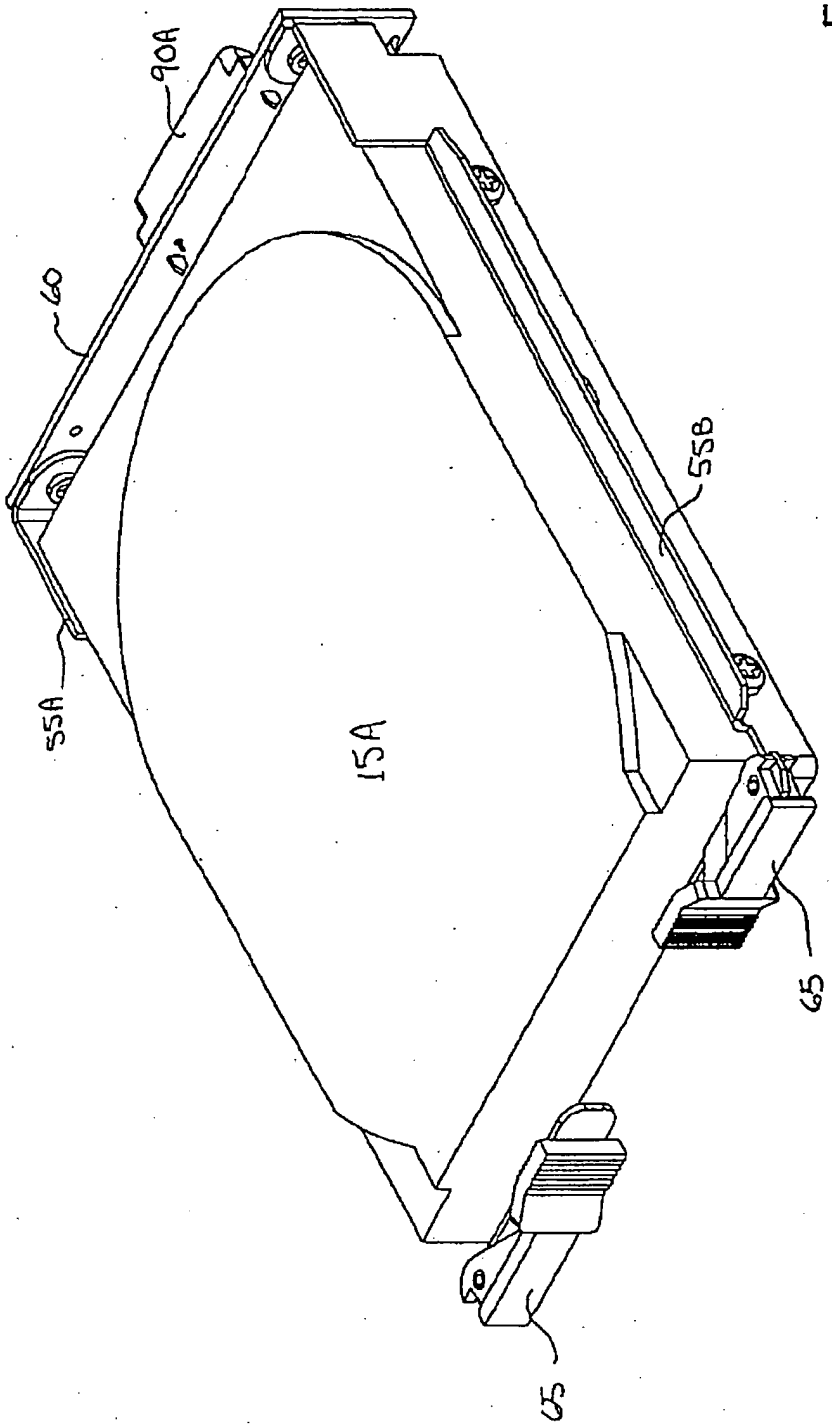
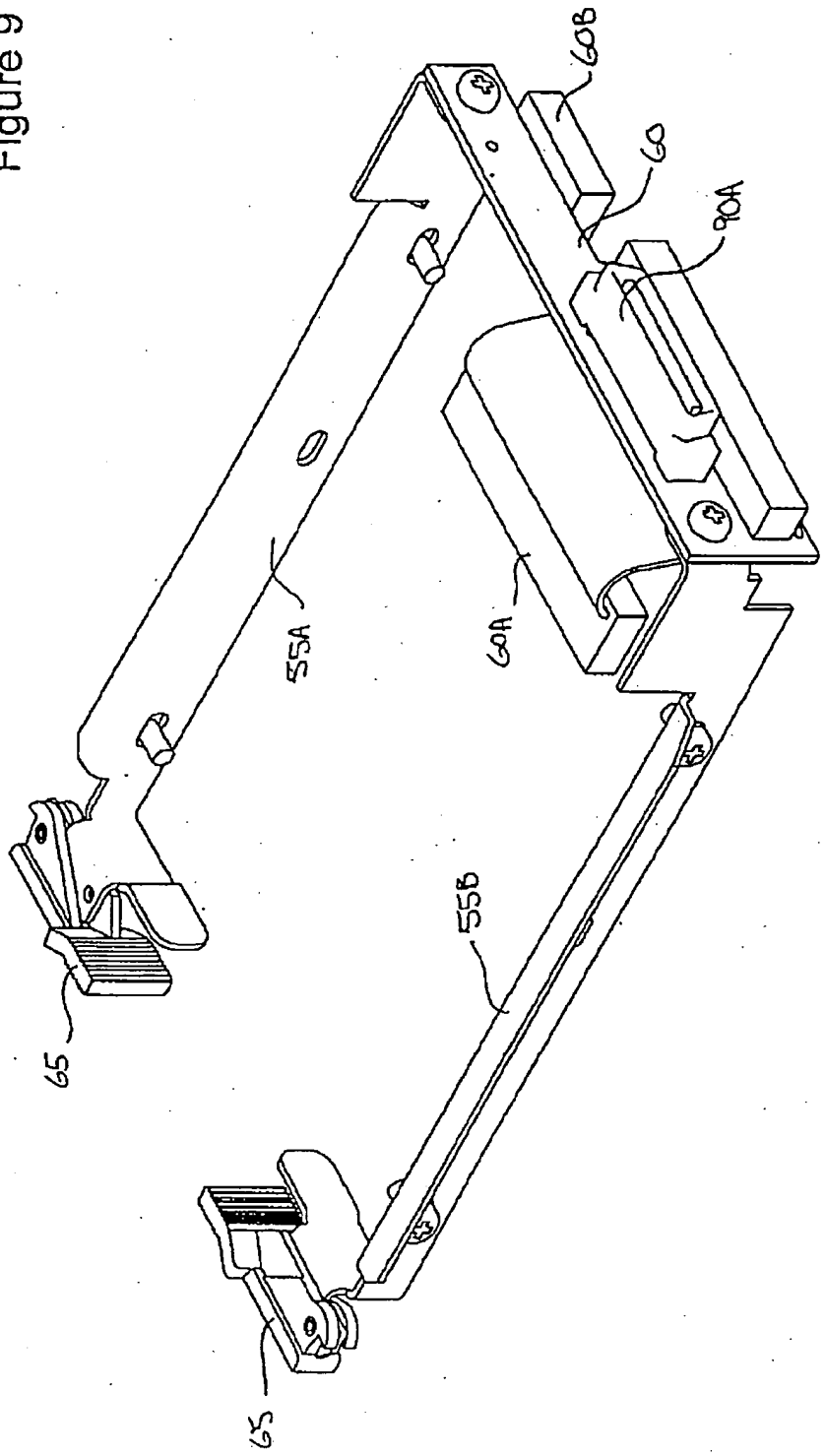


Figure 8

Figure 9



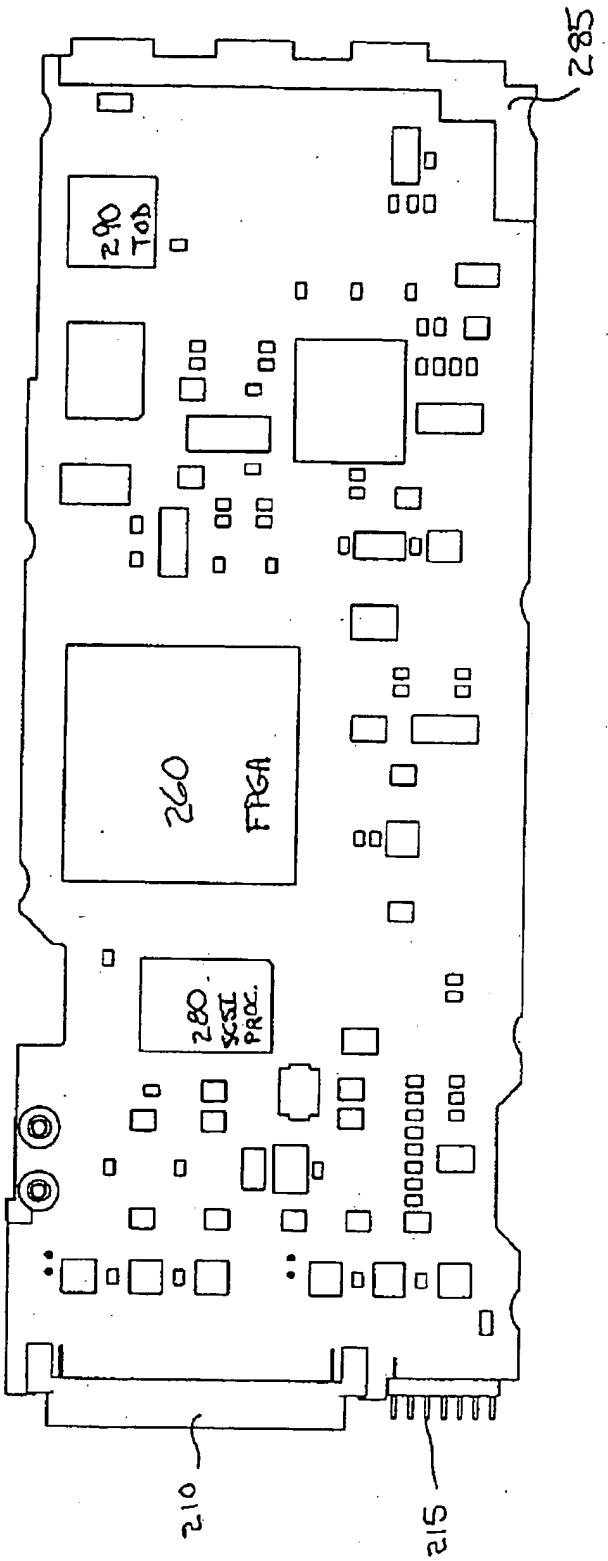


Figure 10A

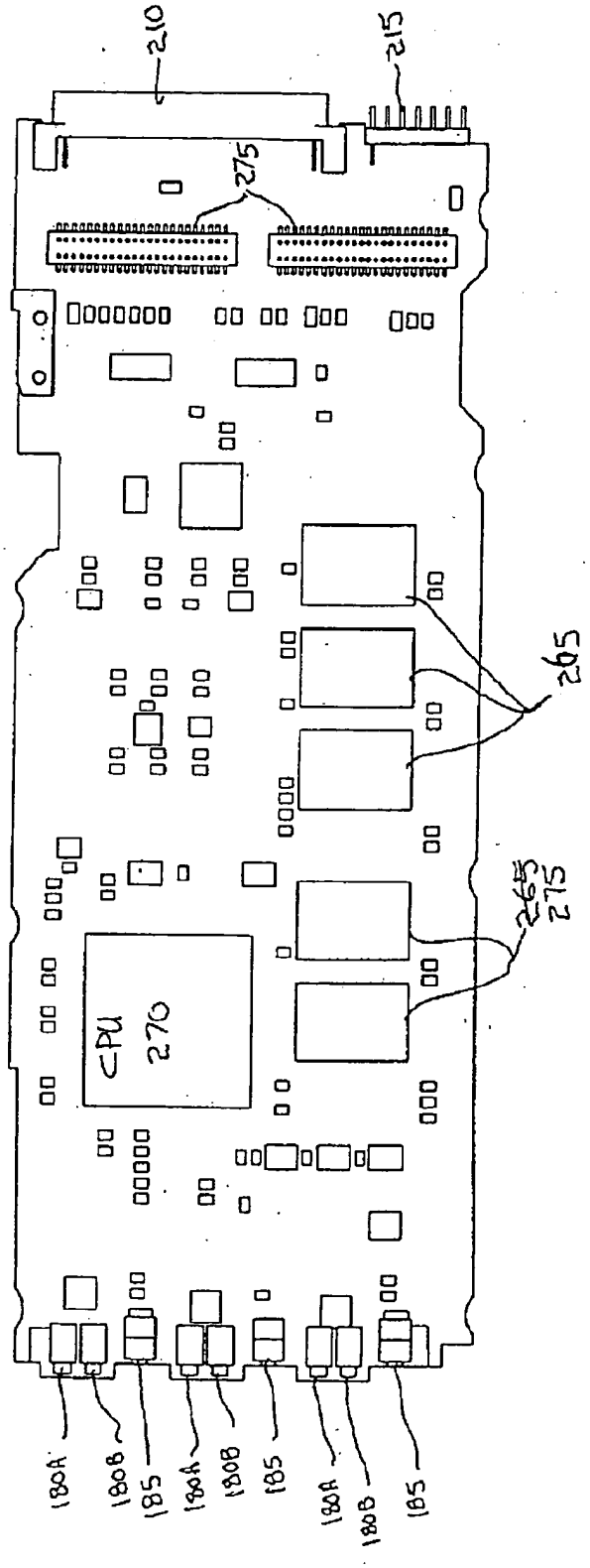


Figure 10B

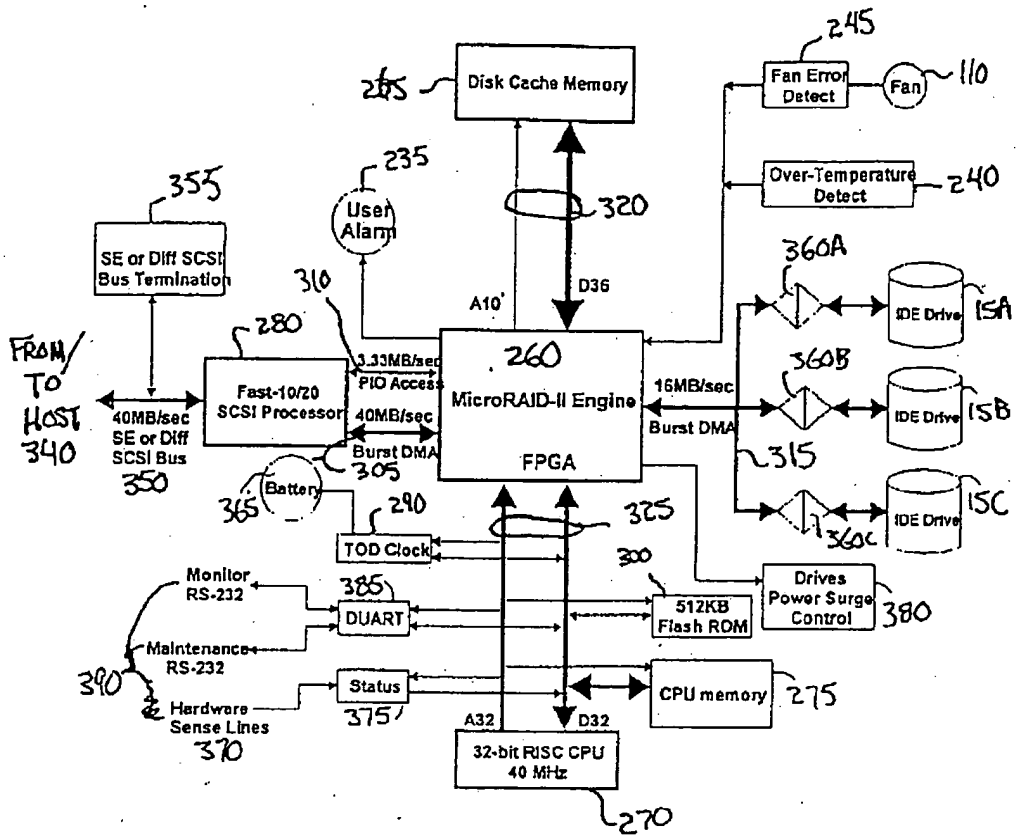


Figure 12

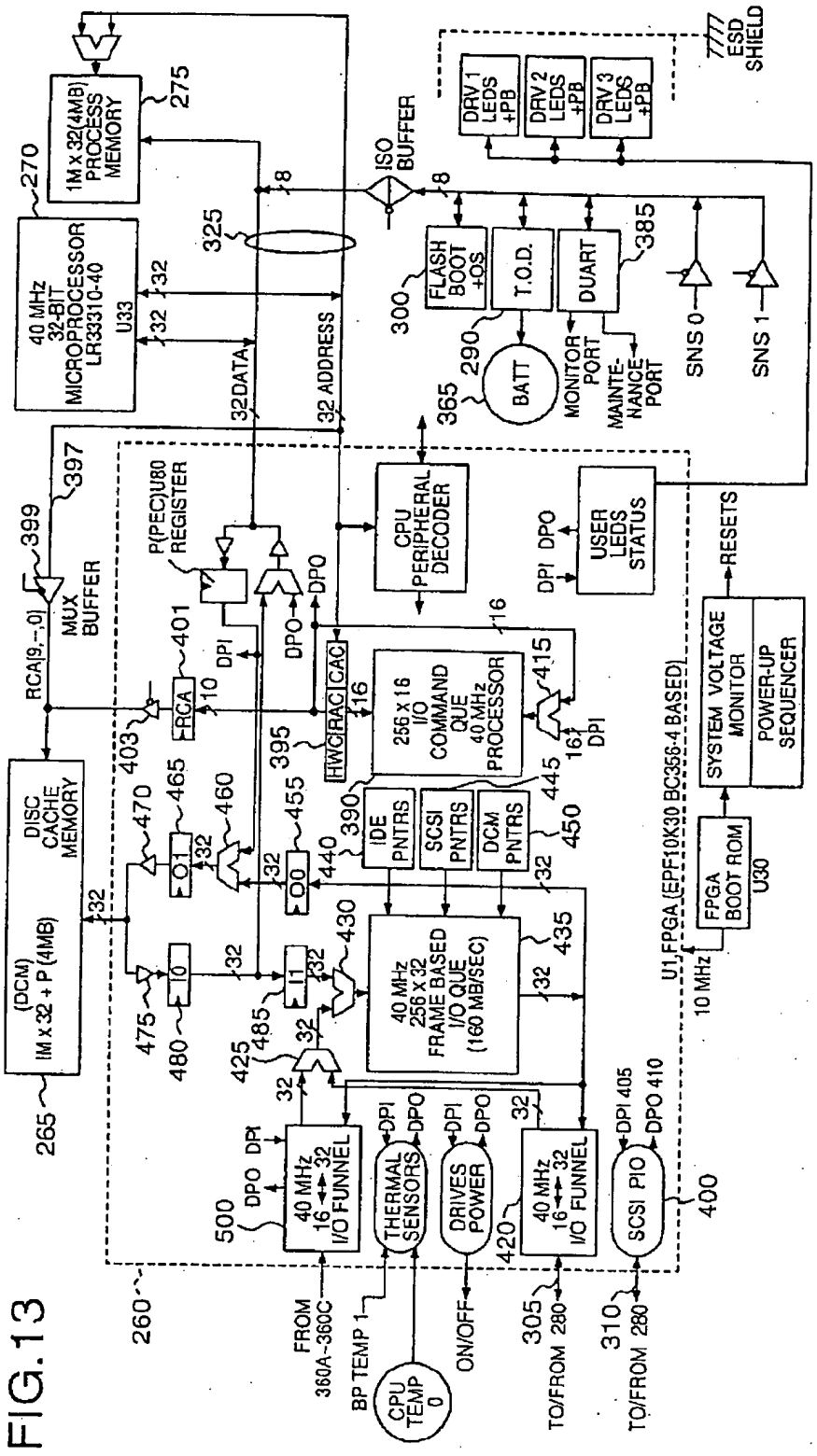


FIG.13

FIG.14

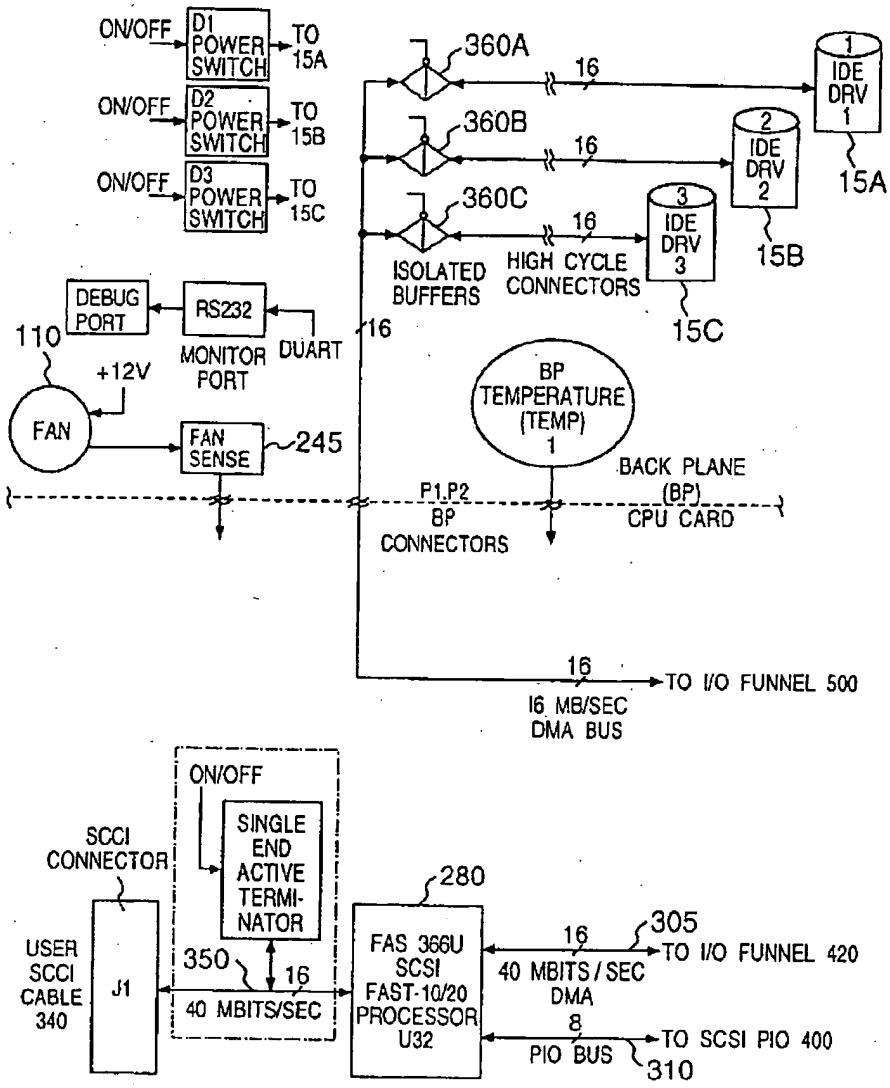


FIG.15

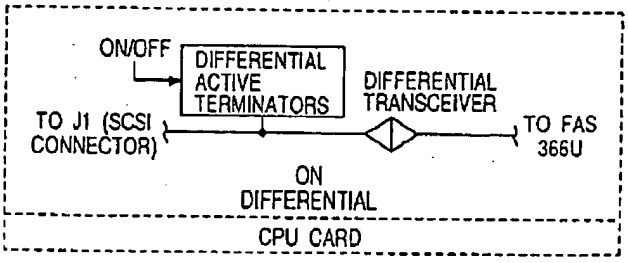
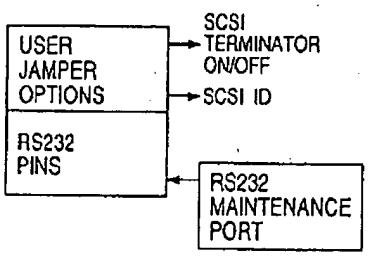


FIG.16



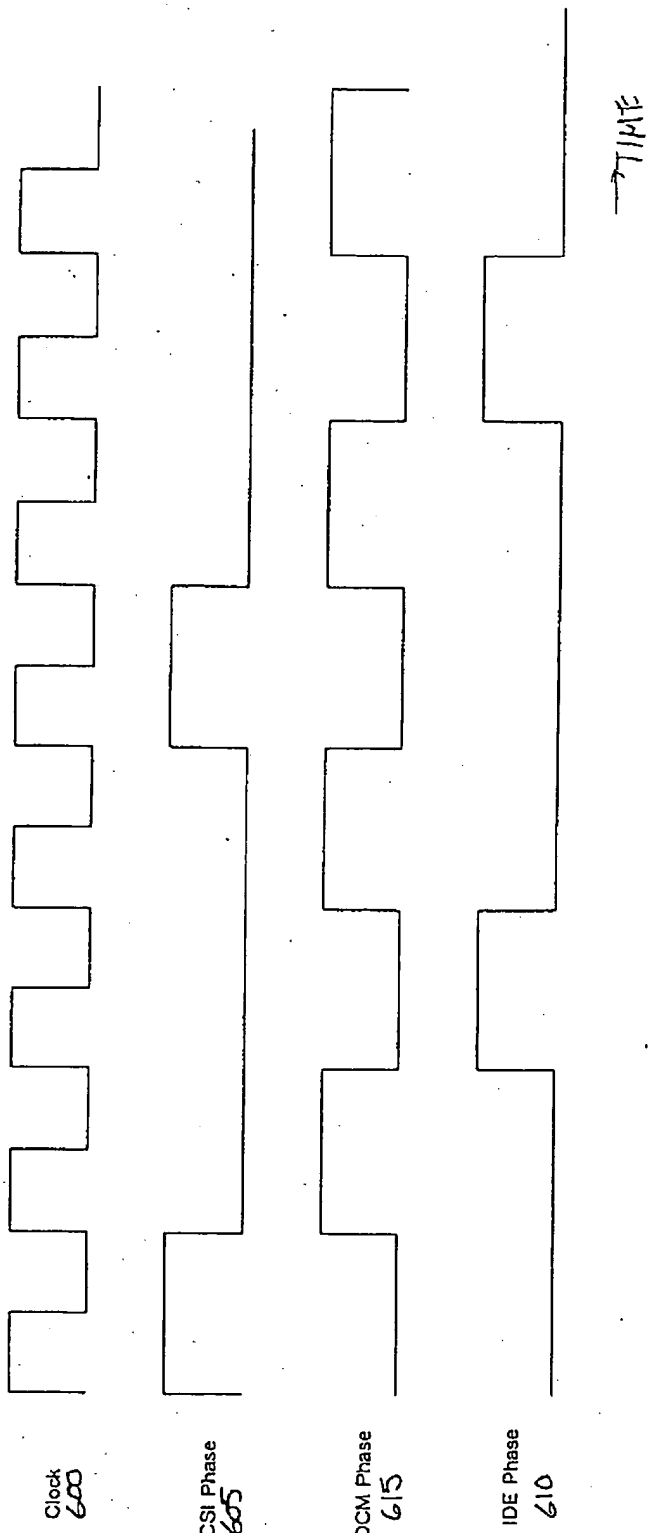
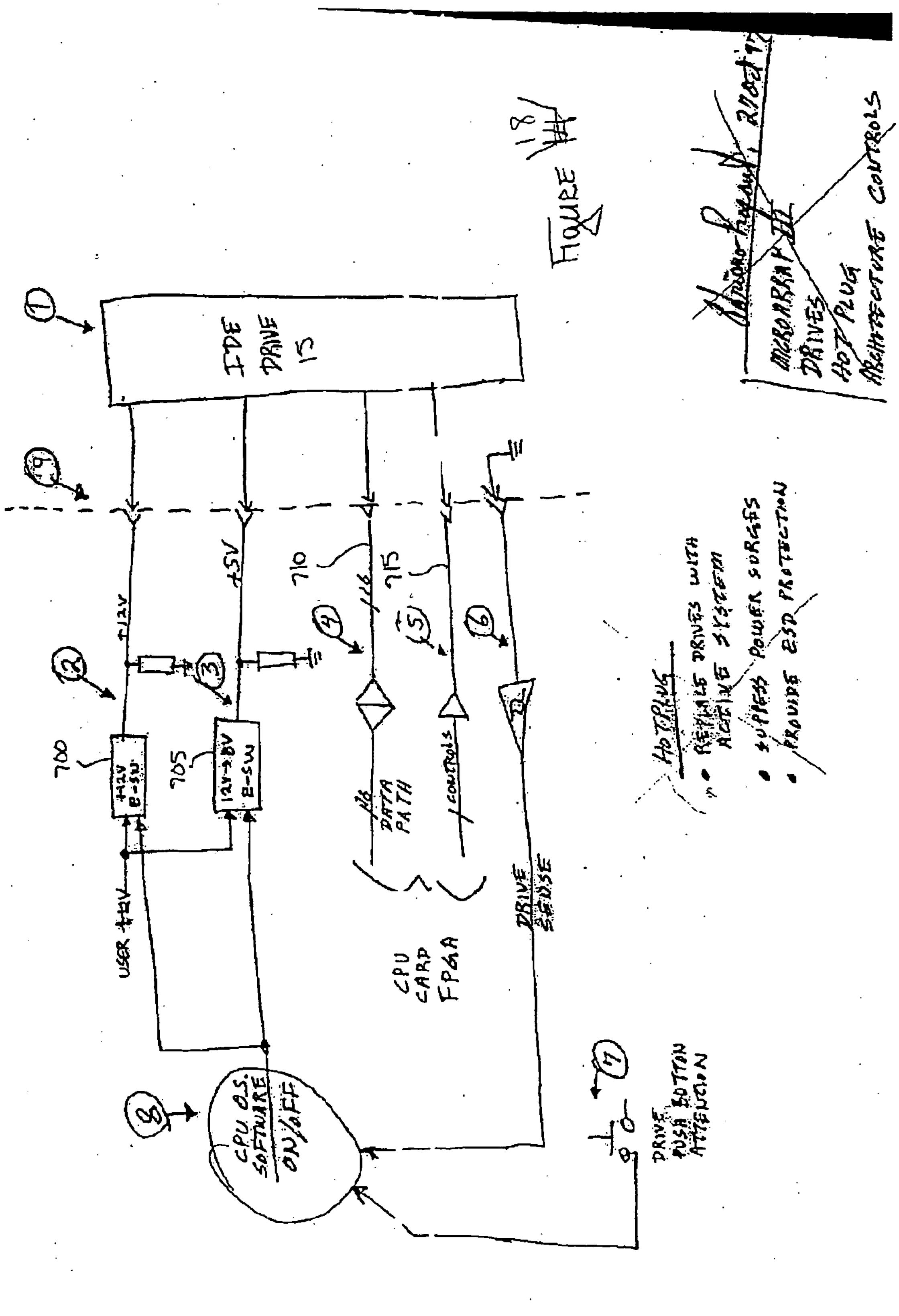


Figure 17



1 Abstract

ABSTRACT OF THE INVENTION

The present invention provides a method and apparatus for a mass storage subsystem such as a RAID array. The invention includes a housing which defines first and second cavities with the first cavity housing an array controller such as a RAID controller. The second cavity houses a plurality of substantially conventional IDE drives conforming to the 3.5" form factor. The array is configured to maximize cooling of the array controller and the drives within the extremely small space defined by the housing.

2 Representative Drawing

Fig. 2